Sveučilište u Zagrebu

FAKULTET ELEKTROTEHNIKE I RAČUNARSTVA

Valentina Zadrija

LOKALIZACIJA OBJEKATA ODOZDO PREMA GORE PRIMJENOM FISHEROVIH VEKTORA

DOKTORSKI RAD

Zagreb, 2017.

Sveučilište u Zagrebu

FAKULTET ELEKTROTEHNIKE I RAČUNARSTVA

Valentina Zadrija

LOKALIZACIJA OBJEKATA ODOZDO PREMA GORE PRIMJENOM FISHEROVIH VEKTORA

DOKTORSKI RAD

Mentor: izv. prof. dr. sc. Siniša Šegvić

Zagreb, 2017.

University of Zagreb

FACULTY OF ELECTRICAL ENGINEERING AND COMPUTING

Valentina Zadrija

BOTTOM-UP OBJECT LOCALIZATION WITH FISHER VECTORS

DOCTORAL THESIS

Supervisor: Associate Professor Siniša Šegvić, PhD

Zagreb, 2017

Doktorski rad izrađen je na Sveučilištu u Zagrebu Fakultetu elektrotehnike i računarstva, na Zavodu za elektroniku, mikroelektroniku, računalne i inteligentne sustave. Istraživanje je potpomognuto od stane Hrvatske zaklade za znanost u okviru projekta *MULTICOLD*, broj I-2433-2014. Voditelj projekta je prof. dr. sc. Siniša Šegvić.

Mentor: izv. prof. dr. sc. Siniša Šegvić

Doktorski rad ima: 134 stranice

Doktorski rad br.: _____

O mentoru

Siniša Šegvić je rođen 1971. godine u Splitu. Osnovnu školu i matematičku gimnaziju završio je u Zadru, osim osmog razreda osnovne škole kojeg je pohađao u Milanu, Italija. Od lipnja 1996. godine do danas, zaposlen je na Zavodu za elektroniku, mikroelektroniku, računalne i inteligentne sustave Fakulteta elektrotehnike i računarstva u Zagrebu. U svibnju 2000. godine obranio je magistarski rad pod naslovom "Uporaba projekcijske geometrije i aktivnog vida u tumačenju scena". Doktorsku disertaciju pod naslovom "Višeagentsko praćenje objekata aktivnim računalnim vidom" obranio je u lipnju 2004. U srpnju 2006. okončao je jednogodišnje postdoktorsko usavršavanje na institutu IRISA u Rennesu, Francuska, na području primjene računalnog vida u samostalnoj navigaciji vozila u urbanom okruženju. U rujnu 2007. okončao je jednogodišnje postdoktorsko usavršavanje na tehničkom sveučilištu u Grazu, Austrija, na području analize nesigurnosti procjene geometrije dvaju pogleda.

Siniša Šegvić je sudjelovao u radu više domaćih i inozemnih znanstvenih projekata. Vodio je jedan istraživački projekt u suradnji s gospodarstvom (HrZZ, 2008-2011), jedan projekt primjene informacijske tehnologije (MZT, 2003-2004), te jedan razvojni projekt Sveučilišta u Zagrebu (2012). Bio je suvoditelj jednog bilateralnog austrijsko-hrvatskog projekta (MZOŠ, 2010-2012). Voditelj je istraživačkog projekta MULTICLOD (HrZZ, 2014-2017). Njegovi znanstveni, istraživački i profesionalni interesi uključuju računalni vid, obradu slike, programsko inženjerstvo te objektno i generičko programiranje. Samostalno odnosno kao koautor objavio je više članaka u časopisima sa međunarodnom recenzijom te na međunarodnim znanstvenim skupovima. Kao recenzent, sudjelovao je u prosudbi članaka podnesenih za objavljivanje na znanstvenim skupovima i u časopisima.

Tijekom rada na Fakultetu elektrotehnike i računarstva, Siniša Šegvić je održavao predavanja na kolegijima Dinamička analiza scena, Oblikovni obrasci u programiranju, Arhitektura i organizacija računala, Skriptni jezici, Inteligentni sustavi, Duboko učenje i Modeli za predstavljanje slike i videa. Za potrebe nastave, priredio je veći broj didaktičkih tekstova, koji su dostupni na mrežnim stranicama fakulteta. Suautor je knjige Python za znatiželjne. Konačno, sudjelovao je i u radu fakultetskog Odbora za istraživanje i međunarodnu suradnju.

Siniša Šegvić vrlo dobro poznaje engleski i talijanski, a služi se i francuskim jezikom. Član je strukovne udruge IEEE. Oženjen je i ima troje djece.

About the Supervisor

Siniša Šegvić was born in 1971 in Split, Croatia. He completed elementary school and high school in Zadar, Croatia, with one year abroad in Milano, Italy (1985-86). He received the BS degree in electrical engineering (9 semesters) in 1996, from the Faculty of Electrical Engine-

ering at the University of Zagreb, Croatia. From 1996 to 2005, he has been employed at the Department of Electronics, Microlelectronics, Computer and Intelligent Systems of the same Faculty, as a teaching assistant. From June 2006, he is employed at the same faculty as an assistant professor.

Siniša Šegvić participated in several national research projects (2 Croatian and 1 french), one Croatian national technology project and was a leader of an another technology project. He is currently in charge of the research project MULTICOLD: Multiclass object detection (HrZZ, 2014-2017). He received MS (2000) and PhD (2004) degrees in computer science from the University of Zagreb, Croatia. In 2005, he started a one-year postdoc position at IRISA, Rennes, France, in the field of appearance-based navigation by monocular computer vision. In 2006, he starts a one-year postdoc position at TU Graz, Austria, in the field of monocular simultaneous localization and mapping, funded by a Marie Curie international incoming fellowship.

His research and professional interests include 3D, active and distributed computer vision, especially in the context of the localization and mapping for navigation purposes. He is also interested in image processing, software engineering, and generic and object oriented programming. He is the author or co-author of several papers published in international conference proceedings and reviewed scientific journals.

Siniša Šegvić speaks english and italian very well, and has basic communication skills in french. He is married and has three children. He is a member of IEEE.

Zahvala

Zahvaljujem se prof. dr.sc. Siniši Šegviću za vodstvo, potporu i razumijevanje tijekom čitavog poslijediplomskog studija. Posebno se zahvaljujem dr.sc. Josipu Krapcu na korisnim savjetima i diskusiji.

Zahvaljujem se kolegama dr. sc. Marku Ševroviću i dr. sc. Mariju Mileru na velikoj pomoći prilikom dobavljanja georeferenciranih video zapisa korištenih u okviru eksperimentalnog vrednovanja ove disertacije.

Zahvaljujem se svima u Mireu, a poglavito Davoru, Andriji, Jasni i Fabiani koji su imali razumijevanja za moje obaveze na poslijediplomskom studiju i bodrili me prije svakog ispita.

Zahvaljujem se svojoj čitavoj obitelji: suprugu Dinku; djeci Petrici, Pepiju i Filipu; mami Katarini, tati Stjepanu, sestri Luciji, bratu Josipu, svekru Petru i svekrvi Branki na potpori.

Zahvaljujem se svom suprugu Dinku koji je uvijek bio uz mene. Bez njega ova disertacija ne bi bila moguća.

Posebno se zahvaljujem svojoj sestri Luciji koja je čitala ovu disertaciju i vjerovala u mene kad ni ja sama nisam.

Naposljetku, zahvaljujem se svojoj mami Katarini koja me odgojila da nikad ne odustajem te njoj posvećujem ovu disertaciju.

Sažetak

U okviru ove disertacije razmatraju se problemi klasifikacije slika i lokalizacije objekata u složenim scenama. Posebna je pažnja usmjerena učenju lokalizacijskih modela uz slabi nadzor budući da se na taj način zaobilazi vremenski zahtjevan proces označavanja lokacija objekata u slikama za učenje. U fazi učenja, dostupne su isključivo oznake prisutnosti objekta u slici, dok se u fazi testiranja zahtjeva predikcija lokacija objekata u vidu opisanih poligona kao i u slučaju učenja pod strogim nadzorom.

Predstavljen je pristup za lokalizaciju objekata temeljen na reprezentaciji Fisherovim vektorima i slabo nadziranom učenju rijetkih lokalizacijskih modela. Predstavljanje slike i slikovnih okana Fisherovim vektorima omogućava primjenu lokalizacijskog modela učenog nad opisnicima cjelokupnih slika za proračun odziva slikovnih okana. Pomoću lokalizacijskih modela rijetkih po komponentama umanjuje se utjecaj prenaučenosti i omogućava učinkovit proračun odziva slikovnog okna. Za poboljšanje lokalizacijske točnosti predložena je primjena metričke normalizacije po komponentama Fisherovog vektora slike. Kako bi se omogućila vremenski efikasna primjena normalizacija u fazi lokalizacije, doprinos slikovnog okna odzivu normalizirane slike određuje se aproksimacijom prvog reda. Naposljetku, budući da Fisherovi vektori ne uzimaju u obzir prostorne odnose okana u slici, predložene su reprezentacije lokalnog prostornog rasporeda slikovnih riječi u vidu prostornih Fisherovih vektora i prostornih histograma. Razvijeni su algoritmi kojima se na temelju slikovnih okana pozitivnog odziva generiraju predikcije lokacija objekata u vidu opisanih poligona.

Provedeno je iscrpno eksperimentalno vrednovanje opisanog pristupa na problemima lokalizacije prometnih znakova i pješačkih prijelaza u složenim prometnim scenama. Pokazano je da se opisani pristup lokalizaciji može primijeniti za potrebe automatizacije digitalnog kartiranja. Za lokalizaciju pješačkih prijelaza, predstavljen je nov skup podataka dobiven polu-automatski na temelju dobrovoljno prikupljenih geopodataka iz OpenStreetMap karte i georeferenciranog videa. Eksperimentalni rezultati pokazuju da predloženi lokalizacijski i reprezentacijski modeli postižu iznimno dobre rezultate unatoč različitim obilježjima traženih objekata i učenju uz slabi nadzor.

Ključne riječi: lokalizacija objekata, Fisherovi vektori, slabo nadzirano učenje, rijetki modeli, model prostornog rasporeda, računalni vid, strojno učenje.

Summary

Bottom-up object localization with Fisher vectors

Billions of images are uploaded on the Internet each year. In order to understand the content of these images, the first step is to understand all objects being depicted. Modern systems for understanding such content make extensive use of machine learning methods designed to detect the presence of objects in images and to recover their locations. These problems are often jointly addressed by applying a localization model at many image locations, and reporting objects where a positive response was obtained. Most successful representatives of this approach employ strong supervision at the training stage, which requires that each training image has to be annotated with accurate object locations. The annotation process is time consuming, error prone and subjective, especially in cases where images contain partially occluded, overlapping or very small objects. In such cases, near to pixel-level annotation accuracy may be required for best results, while in realistic scenarios thousands of annotations may be necessary to achieve top performance.

On the other hand, most of the visual content available on social networks like Instagram or Twitter is only partially labeled with so called *hashtags*. These *hashtags* convey a simple description of the image content in a form of a single word, abbreviation or word concatenations prefixed by the # symbol. For instance, *#bus* denotes the presence of a bus object in an image. In order to make use of image-wide labels, many recent approaches attempt to solve the localization problem in a weakly-supervised manner. In this setting, the training procedure is supposed to learn the localization model without having known object locations. At the test time, however, bounding boxes have to be predicted for each learned object class as in the strongly supervised case. This can be useful even if the recovered object classifier is not particularly fast, since the recovered object locations can be used to train a more efficient localization model in a strongly supervised fashion.

Following the described reasoning, as the first contribution of this thesis, we have proposed a novel weakly-supervised object localization method based on Fisher vectors (FV) and model sparsity at the component level. The method relies on the Fisher embedding of local features (e.g. convolutional features from a deep neural network, SIFT) for representing the image

patches and entire images. The Fisher vectors extend the well known Bag of Visual Words (BoVW) representation and are chosen for the task because of many desirable properties. Most notably, the FV representation attenuates the background information, preserves and enhances unusual details, which is crucial for the classification in weakly supervised setting. The second desirable property of Fisher vectors is additivity. Given the assumption that the patches are independently and identically distributed, the Fisher vector of an entire image can be obtained by simple average pooling. Moreover, due to their nonlinearity, the Fisher vectors are amenable for classification with linear classifiers. Linear localization models are efficient to evaluate and learn (linear in the number of training samples). Hence, by using the Fisher vectors of entire images accompanied with image-wide labels, a competitive linear model can be trained. At the training time, we enforce the model sparsity on a component level, where a particular component of the Fisher vector can be seen as a visual word. In this way, all coefficients corresponding to non-informative visual words are set to zero and a small subset of discriminative visual words is identified. The localization is then performed in a bottom-up fashion, i.e. the obtained localization model is applied in a sliding window manner to image patches. In order to make the localization efficient, we have proposed two fold optimizations which exploit the model sparsity. First, image patches, which are not assigned to the selected set of visual words, are discarded. This significantly reduces the number of patches, similar to the effects of the first few stages of a cascaded classifier. Second, for the remaining patches only the parts of the Fisher vector which correspond to discriminative visual words are computed. Consequently, the patch score with the model is computed only for a fraction of the high dimensional representation making the localization efficient. From the patches with a positive classification score, object predictions are created. For that purpose two novel algorithms are proposed: i) the algorithm based on independent processing of top rated patches on each scale, and ii) the algorithm focused on combining per-scale responses in a unified heat map. Both algorithms consider a subset of patches with the highest score to generate bounding polygon predictions. The size of the subset is determined as a parameter of the algorithm. In order to cope with possible multiple localization responses, a non-maximum suppression algorithm is proposed.

In order to improve the localization accuracy, we have proposed two methodological novelties focusing on the application of the non-linear Fisher vector normalizations. The non-linear normalizations include power and metric normalizations which are applied to Fisher vectors of entire images before training the localization model. In order to exploit the Fisher vector structure, as a second contribution of this thesis, we have proposed using the intra-component metric normalization in conjunction with the component-level sparsity. In this setting, the metric normalization is separately applied to the parts of the Fisher vector corresponding to different visual words. The intra-component metric normalization accounts for the so called “burstiness phenomenon”, where certain visual word can be unusually frequent in a particular image. The

experiments have shown that this setting further increases the model sparsity and reduces the localization miss frequency. On the other hand, the non-linear normalizations invalidate additivity of the Fisher vector representation. In such setting, the linear decomposition of an image score into a sum of patch scores is no longer possible. The patch contribution to the image score can be computed directly by: i) subtracting each patch FV from an image representation, ii) normalizing the obtained results, iii) scoring them with the model and iv) subtracting the obtained scores from the normalized image score. Unfortunately, this procedure is computationally very complex because it involves expensive square rooting and power operations per each image patch. This makes its application to all image patches impractical. Hence, as a third contribution of this thesis, we have proposed a first-order approximation which corresponds to taking the dot-product between the un-normalized patch representation and the gradient of the normalized image score. In this way the normalizations are computed only once, when computing this gradient vector. The experiments have shown that this approximation achieves comparable localization accuracy with respect to the direct approach, while gaining a huge overall execution speedup.

The so far described weakly supervised localization pipeline disregards the spatial relations between image patches. The experiments have shown that such framework successfully identifies patches responsible for the image label. Still, some of the difficult background patches may be scored positively and as such generate false alarms. The analysis of positive responses has revealed that the spatial relation between the patches assigned to particular visual words in the background differ from those on the object of interest. Therefore, as the fourth contribution of this thesis, we have proposed two spatial descriptors to capture these relations, i.e. spatial histograms and spatial Fisher vectors. These spatial representations are formulated as second-level descriptors, which describe the pairwise spatial layout in a local neighbourhood around positive responses. By using the spatial descriptors, a sparse second-level localization model is learned. The experiments have shown that the obtained model filters out the difficult background patches and improves the localization accuracy. Finally, object predictions are created from all patches scored positively by the second level model. In this way, the number of parameters of the algorithm responsible for creating the object predictions is also reduced.

The extensive experimental evaluation of the proposed contribution was performed on two real-world problems: i) automating the road safety inspections and ii) automating the road-environment mapping in order to enrich cartographic data. For the task of automating the road safety inspections, an adapted version of a public traffic sign dataset was used. The traffic signs were chosen as a representative example of the road infrastructure which covers less than one percent of the image area in the target dataset. As such, the localization of the traffic signs presents an interesting challenge in a weakly supervised setting. The experiments have shown that the appearance based pipeline built upon SIFT features successfully localizes such objects

with the average precision of 86 percent and miss frequency of 0.12 (12 percent of objects had remained undetected). Furthermore, by using the proposed spatial descriptors, localization precision has improved by 4 pp, while the miss frequency has dropped by 5 percentage points.

For the task automating the road-environment mapping, a novel dataset containing 2381 images of pedestrian crossings (zebra crossings) was introduced. The dataset was obtained by semi-automated matching of OpenStreetMap data (longitude, latitude) to GPS references of video frames. The dataset is characterized by a large intra-class variation and contains various background objects (mainly road surface markings) with very similar patterns as the object of interest. These "co-occuring distractors" make the localization in the weakly supervised setting extremely difficult. Given the nature of the dataset, a permissive threshold of IoU > 0.10 was used for the purpose of evaluation (for traffic signs standard value of IoU > 0.50 was used). By using the appearance-based framework on top of convolutional features from the VGG-E network, average precision of 92 percent accompanied with a miss frequency of 0.25 was obtained. The analysis of failure cases has shown that most of the localization errors are a result of either i) distant objects which appear very small, or ii) multiple nearby objects which get reported as one object. By increasing the evaluation threshold to IoU > 0.30, an accuracy of 80 percent and miss frequency of 0.30 was obtained. We believe that this is still a fairly good result due to the nature of the dataset and weak supervision.

The experiments on both datasets have shown several interesting points that confirm the significance of the proposed contributions. First, component-sparse models outperform widely used ℓ_2-dense and ℓ_1-sparse models. The advantage of the component sparsity is especially significant in comparison to the ℓ_2-dense models (16 pp for pedestrian crossings, 21 pp for traffic signs). At the same time, the component sparsity results in a huge gain in the execution speed since the component-sparse models use only a fraction of image representation (5 percent for pedestrian crossings, 1 percent for traffic signs). Second, the intra-component metric normalization in conjunction with the component-level sparsity further increases the model sparsity and reduces the localization miss frequency. Finally, the first-order approximation of the normalized FV score achieves comparable localization accuracy with respect to the direct approach, while gaining a huge overall execution speedup (around 200 times on the traffic sign dataset).

To conclude, the experimental evaluation has shown that general purpose object localization models can be trained from geo-referenced video and crowd-sourced image-wide labels provided by services such as OpenStreetMap. The obtained location information can be used to verify correctness of the road infrastructure, e.g. the placement of a freshly painted pedestrian crossing or the appearance of a worn-out traffic sign. The proposed method is also applicable to a variety of other real-world tasks, especially since much of the online visual content is labeled with weak labels. The recovered localization models could be used to inspect city assets such as park benches, waste disposal baskets or trees and recognize types of clothes or shoes in fashion

show images.

Several interesting directions for future work are possible. The experiments have shown that the non-linear normalizations greatly improve the localization performance for the appearance-only models. Hence, one interesting direction would be to apply the non-linear normalizations for localization with spatial models. Furthermore, the proposed method is compatible with different low level descriptors, i.e. we used SIFT for traffic sign and convolutional features for pedestrian crossing localization. For future work, we intend to perform additional experiments with these low level features on both datasets. We believe that with these new experiments we might gain further insight about choosing the most suitable low level features for the given task. Another interesting direction would be to fully automate the weakly supervised training from crowd-sourced GPS labels. Due to the fact that the OpenStreetMap data is obtained by different users with different GPS devices, 15 percent of the automatically generated images did not contain pedestrian crossings. Hence, an additional step of manual pruning was performed in order to account for the noise in the data. This manual step of filtering could be sidestepped by learning the localization model directly from the noisy crowd-sourced GPS labels. Finally, due to the recent success of convolutional neural networks in various areas of computer vision, a suitable direction for future work includes *end-to-end* learning for the task of weakly supervised localization.

Keywords: object localization, Fisher vectors, weakly supervised learning, sparse localization models, spatial layout representations, computer vision, machine learning.

Sadržaj

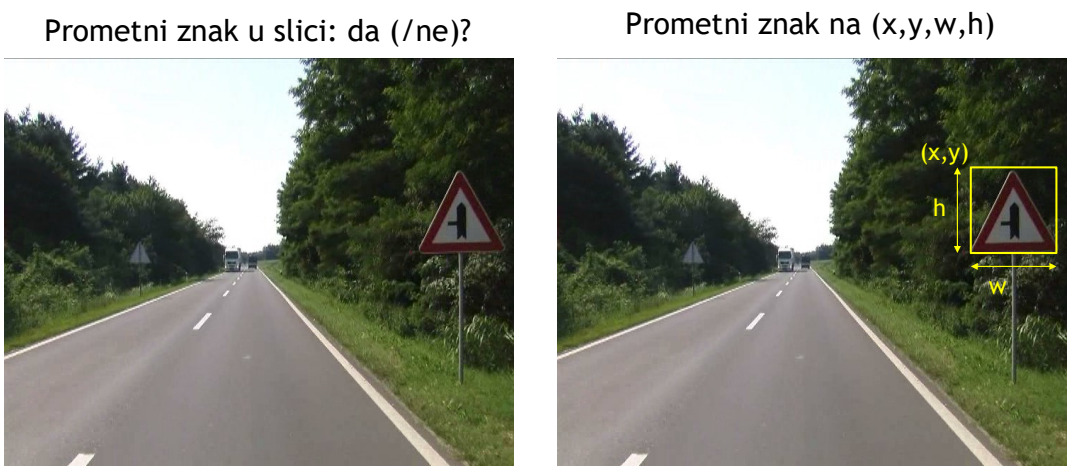
Poglavlje 1

Uvod

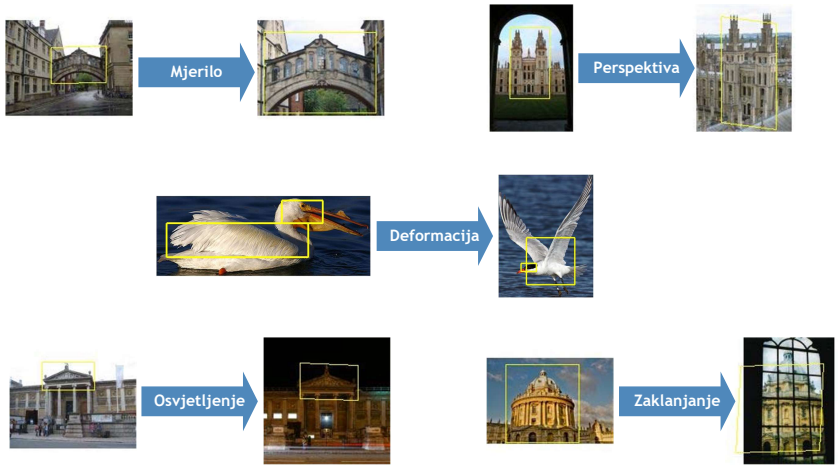
Računalni vid obuhvaća multidisciplinarno područje koje se bavi strojnom obradom slikovnog sadržaja. Jedan od temeljnih zadataka računalnog vida jest određivanje prisutnosti i lokacije objekata u slikama. Navedeni problemi ilustrirani su u okviru slike 1.1 na primjeru objekta razreda prometnih znakova. Problem određivanja prisutnosti objekta u slici naziva se još i problemom klasifikacije slika. Cilj klasifikacije slika jest odrediti razrede (klase) objekata relevantne za sliku. Neke od primjena klasifikacije slika uključuju organizaciju slika u semantički smislene kategorije ili pretraživanje slika na temelju tekstualnih upita (engl. *query by text image retrieval*). Zadatak lokalizacije objekata odnosi se na problem određivanja opisanog poligona traženog objekta (engl. *bounding polygon*). Određivanje egzaktne lokacije objekta u složenim scenama iznimno je težak problem iz niza razloga. Glavni problemi i izazovi ilustrirani su na slici 1.2 te uključuju:

- promjene u perspektivi (očištu i gledištu kamere) (engl. *viewpoint changes*): promatrani objekt može biti rotiran u ravnini (engl. *in-plane rotation*) ili izvan ravnine (engl. *out-of-plane rotation*)
- promjene u mjerilu (engl. *scale*): lokalizacijski model treba biti u stanju utvrditi lokacije jako malenih i velikih objekata
- fotometrijske transformacije, odnosno promjene u osvjetljenju scene (engl. *illumination changes*)
- djelomično zaklonjene objekte (engl. *occluded objects*)
- objekte elastične strukture (engl. *deformable objects*): primjer prikazan na slici 1.2 ilustrira objekte razreda ptica, gdje ptica može biti uhvaćena u letu raširenih krila ili pak sklopljenih krila.

Najbolji rezultati (engl. *state of the art*) u rješavanju opisanog problema postižu se metodama utemeljenim na strogo nadziranom učenju [3, 4, 5, 6, 7]. Strogo nadzirano učenje lokalizacijskih modela (engl. *strongly supervised learning*) zahtijeva označavanje lokacija objekata u vidu opisanih poligona (engl. *bounding polygon, bounding box*). Proces označavanja je du-



Slika 1.1: Ilustracija zadataka klasifikacije slika (lijevo) i lokalizacije objekata (desno). Prometni znak upozorenja predstavlja traženi objekt.

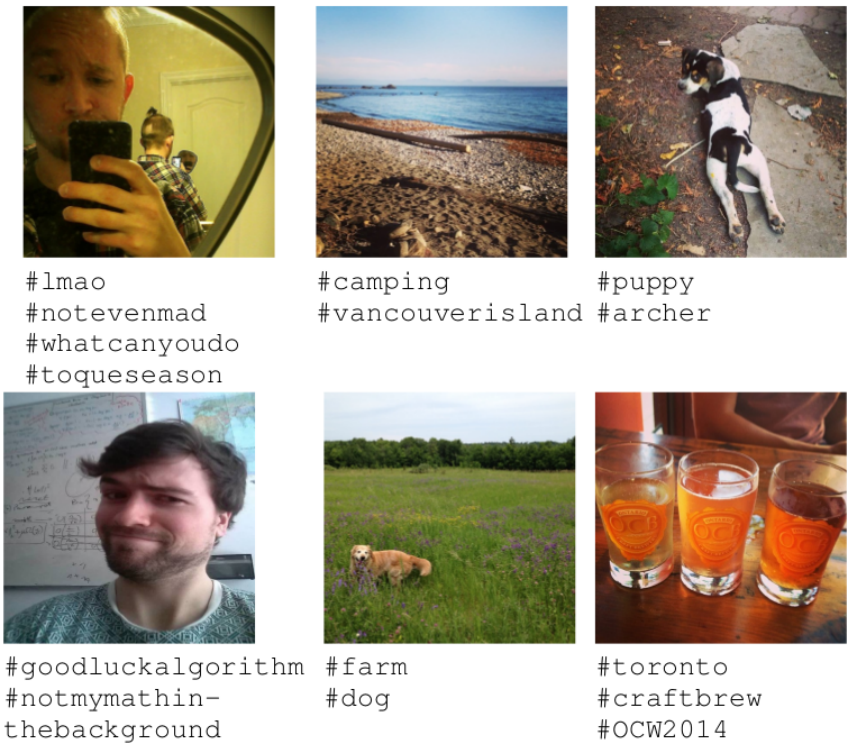


Slika 1.2: Motivacija: problemi i izazovi detekcije prisutnosti i lokalizacije objekata. Prilagođeno prema [1, 2].

gotrajan, subjektivan i podložan pogreškama. Ovisno o složenosti objekta traženog razreda, često je potrebno pribaviti i nekoliko tisuća oznaka kako bi se postigli zadovoljavajući rezultati. Označavanje malenih objekata kao što su prometni znakovi zahtijeva da oznake objekata budu točne na razini piksela. S druge strane, većina slikovnog sadržaja dostupnog na Internetu tek je djelomično označena [8]. Svakodnevno se na poslužitelje različitih servisa (primjerice, Instagram, Pinterest, Twitter, Facebook) pohranjuju goleme količine slika označenih primjenom kratkih oznaka (engl. *hashtag*) koje ukratko opisuju sadržaj slike. Slika 1.3 ilustrira različite primjere primjere takvih oznaka dane uz slike ljudskih lica i odgovarajućih emocija, znamenitosti neživе prirode, životinja te objekata karakterističnih za scene interijera (engl. *indoor objects*).

U literaturi se može pronaći više paradigmi strojnog učenja kojima se nastoji zaobići problem prikupljanja oznaka lokacije objekata:

- nenadzirano učenje (engl. *unsupervised learning*) [11, 12]



Slika 1.3: Primjer „slabih" oznaka (engl. *hashtag*) na razini slika [9, 10].

- polu-nadzirano učenje (engl. *semi-supervised learning*) [13, 14]
- aktivno učenje (engl. *active learning*) [15, 16, 17]
- slabo nadzirano učenje (engl. *weakly supervised learning*) [18, 19, 20, 21].

Nenadzirano učenje Kod nenadziranog učenja dan je skup podataka bez ikakvih oznaka pripadnosti određenom razredu. Cilj je pronaći pravilnosti unutar podataka, odnosno strukturu podataka. Primjeri nenadziranog učenja uključuju grupiranje (engl. *clustering*), procjenu funkcije gustoće vjerojatnosti prema kojoj su podaci generirani (engl. *density estimation*) i postupke smanjenja dimenzionalnosti podataka (engl. *dimensionality reduction*).

Polunadzirano učenje Za učenje lokalizacijskog modela polunadziranim učenjem koristi se skup podataka sačinjen od veće količine neoznačenih podataka i manjeg udjela označenih. Učenjem se istovremeno optimira klasifikacijski gubitak nad označenim podacima i rekonstrukcijski gubitak nad neoznačenim podacima.

Aktivno učenje Aktivno učenje označava iterativan proces učenja u okviru kojeg algoritam odabire primjere za učenje na način da postavlja upite učitelju (najčešće ljudskom agentu) u obliku neoznačenih instanci koje je potrebno označiti.

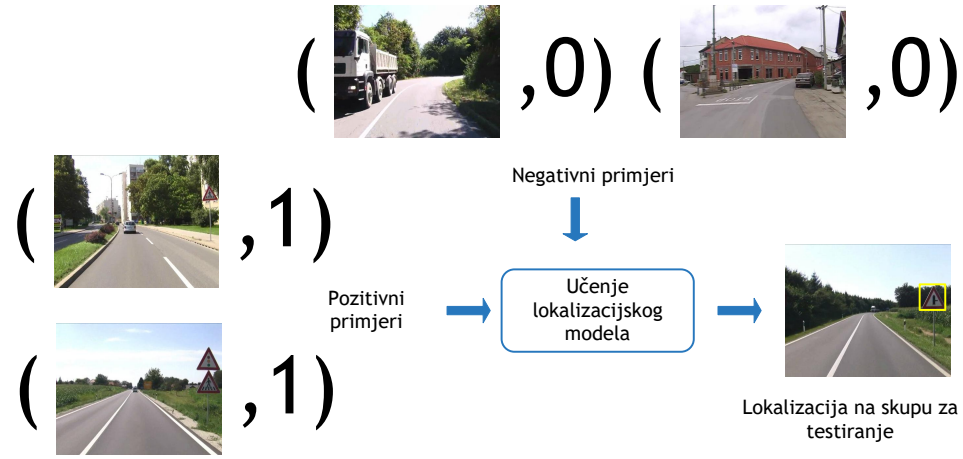
Slabo nadzirano učenje Kod slabo nadziranog učenja, nadzor uključuje oznake na razini slike. Cilj učenja je odrediti lokalizacijski model isključivo na temelju informacije nalazi li se u slici objekt pojedinog razreda ili ne.

U okviru ove disertacije razmatra se lokalizacija objekata temeljena na postupcima slabo nadziranog učenja.

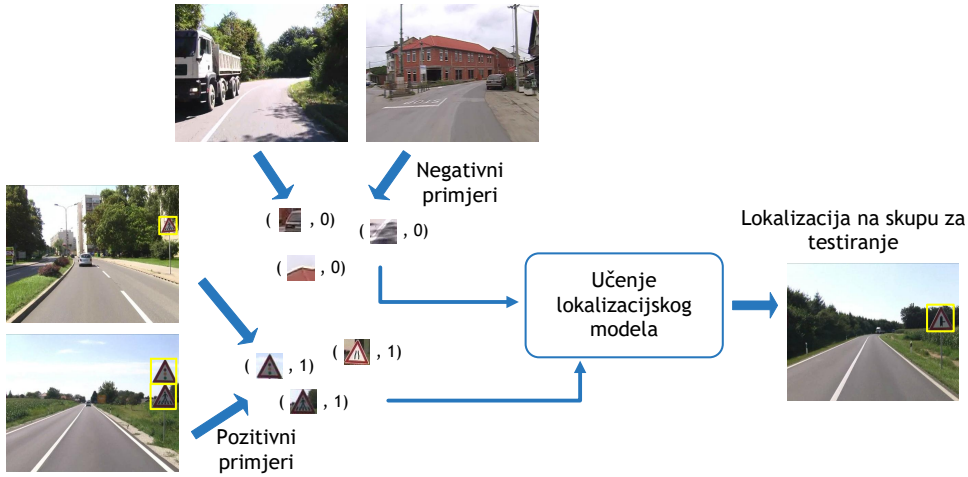
1.1 Opis problema

Osnovne značajke slabo nadzirane lokalizacije i razlike u odnosu na strogo nadzirani pristup ilustrirane su u okviru slike 1.4 na primjeru lokalizacije prometnih znakova. Ulaz u algoritam slabo nadziranog učenja (gornji redak u slici 1.4) čine pozitivne slike za koje postoji informacija da sadrže objekt traženog razreda i negativne slike koje ne sadrže tražene objekte. Zadatak algoritma učenja jest otkriti lokacije objekata u slikama za učenje i na temelju njih naučiti lokalizacijski model koji na još neviđenim slikama uspješno generira predikcije lokacija objekata u vidu opisanih poligona baš kao i u strogo nadziranom slučaju (donji redak). Uz pretpostavku da svaka pozitivna slika sadrži samo jedan objekt (što nije slučaj u okviru eksperimenata predstavljenih u okviru ove disertacije), iscrpno pretraživanje prostora mogućih lokacija objekata rezultira eksponencijalnom složenosti W^T, gdje W odgovara broju slikovnih regija, a T broju slika u skupu za učenje. Računalnu složenost učenja moguće je umanjiti postupcima uzorkovanja [22], grupiranja [23] ili predviđanjem potencijalnih lokacija objekata dobivenih segmentacijom [18, 24] ili naučenom mjerom *objektnosti* [25, 26, 27, 28, 29]. Nedostatak predloženih pristupa leži u činjenici da se lokacije pojedinih objekata mogu izostaviti već u fazi odabira potencijalnih lokacija. To posebno može biti problem u slučaju kada objekt traženog razreda zauzima razmjerno malu površinu u okviru slike. Problemu slabo nadziranog učenja može se pristupiti i na način da se lokalizacijski model inicijalno uči nad cjelokupnim slikama [30] ili slikovnim regijama [24], koje se postupno sužavaju ne bi li se naposljetku identificirale lokacije objekata. Za predstavljanje slika ili slikovnih regija u recentnim se radovima koriste histogrami slikovnih riječi (engl. *Bag of Visual Words, BoVW*) [24, 30] ili Fisherovi vektori [18].

U okviru ove disertacije razvijen je postupak slabo nadzirane lokalizacije objekata temeljen na reprezentaciji Fisherovim vektorima [31, 32, 33] i po komponentama rijetkim lokalizacijskim modelima (engl. *group sparse models, block sparse models*) [34, 35, 36]. Pristup pripada skupini postupaka „odozdo prema gore" jer polazi od primitivnih slikovnih elemenata kao što su kvadratna ili pravokutna slikovna okna te na temelju njih gradi složenije slikovne strukture. Slikovna se okna predstavljaju opisnicima niske razine (lokalnim opisnicima). U okviru ove disertacije provedeni su eksperimenti nad dva tipa lokalnih opisnika, a to su SIFT (Scale Invariant Feature Transform) [37] i konvolucijske značajke [38] pete konvolucijske grupe duboke mreže VGG-E [39].



(a) Slabo nadzirana lokalizacija.



(b) Strogo nadzirana lokalizacija.

Slika 1.4: Lokalizacija prometnih znakova: (a) lokalizacijski model dobiven je postupcima slabo nadziranog učenja na temelju informacije o prisutnosti prometnog znaka u slici, (b) lokalizacijski model dobiven je postupcima strogo nadziranog učenja na temelju opisanih poligona prometnih znakova u pozitivnim slikama. U fazi testiranja, u oba se slučaja zahtijeva predikcija lokacija objekata u vidu opisanih poligona.

Lokalni se opisnici transformiraju u prostor visokodimenzionalnih Fisherovih vektora [31, 33]. Fisherovi vektori predstavljaju proširenje histograma slikovnih riječi (engl. *Bag of Visual Words, BoWV*) [40, 41], gdje se kao slikovni rječnik koristi generativni model. U okviru ove disertacije, kao generativni model koristi se model mješavine Gaussovih raspodjela (engl. *Gaussian mixture model, GMM*) [34, 42], gdje pojedina Gaussova komponenta odgovara slikovnoj riječi. Transformacija u prostor Fisherovih vektora obavlja se pomoću nelinearne Fisherove jezgre [43] u odnosu na komponente Gaussova modela. Iz tog razloga, Fisherov vektor odlikuje specifična blokovska struktura, gdje pojedini blok odgovara doprinosu određene Gaussove komponente. Uz pretpostavku da su slikovna okna nezavisno i jednoliko raspodijeljena, Fisherov vektor slike dobiva se jednostavnom operacijom usrednjavanja (engl. *sum pooling*) Fisherovih vektora slikovnih okana [33]. Povoljno svojstvo Fisherovih vektora jest da poništavaju utje-

caj pozadinske informacije u slici i naglašavaju neuobičajene detalje karaketristične za traženi razred objekata. Dodatno, s obzirom na nelinearnost Fisherove jezgre, uzorci u prostoru Fisherovih vektora pogodni su za klasifikaciju linearnim modelima. Na temelju Fisherovih vektora slika uči se linearan lokalizacijski model koji se uslijed aditivnosti Fisherovih vektora okana može primijeniti za proračun odziva okana. Prednost ovakvog pristupa jest da se u fazi učenja zaobilazi postupak sužavanja prostora potencijalnih hipoteza metodama segmentacije [44] i *objektnosti* [26], što ovu metodu čini pogodnom za slabo nadziranu lokalizaciju malenih objekata. Lokalizacijski se model primjenjuje na slikovna okna, a na temelju okana pozitivnog odziva formiraju se predikcije lokalizacijskih poligona.

Vremenska učinkovitost postupka lokalizacije postiže se primjenom klasifikacijskih modela rijetkih po komponentama [34, 35, 36]. Modeli rijetki po komponentama dobivaju se učenjem uz primjenu $\ell_{2,1}$ regularizacijske funkcije koja uzima u obzir blokovsku strukturu Fisherova vektora, gdje se unutar doprinosa pojedine komponente (slikovne riječi) primjenjuje ℓ_2 regularizacija, a među komponentama ℓ_1. Na taj se način obavlja probir diskriminativnih Gaussovih komponenti u okviru algoritma učenja, gdje se svi koeficijenti modela koji odgovaraju nekoj neinformativnoj komponenti postavljaju na vrijednost ništice. Dodatno, smanjuje se utjecaj efekta prenaučenosti* (engl. *overfitting*) [34], čime se osigurava da model dobro generalizira na skupu za testiranje. Na temelju odabranih komponenti, obavlja se filtriranje slikovnih okana, gdje se iz razmatranja uklanjaju slikovna okna dodijeljena neinformativnim komponentama. Na taj se način postiže značajno ubrzanje jer se lokalizacijski model primjenjuje na malom podskupu slikovnih okana. Za preostala slikovna okna, lokalizacijski odziv se evaluira tek za djelić visokodimenzionalne reprezentacije koji odgovara diskriminativnim slikovnim riječima čime se postiže dodatno ubrzanje.

Razvijena su dva algoritma kojima se na temelju okana pozitivnog odziva formiraju lokalizacijski poligoni: algoritam temeljen na pojedinačnom mjerilu i algoritam temeljen na ujedinjenim mapa odziva preko svih mjerila. Algoritam zasnovan na pojedinačnom mjerilu na temelju top T okana najvećeg odziva gradi prostorni graf povezanosti za svako mjerilo. Lokalizacijski se poligoni formiraju na temelju povezanih komponenti grafa. Algoritam je posebno pogodan za lokalizaciju malih objekata te se primjenjuje prilikom lokalizacije prometnih znakova u odjeljku 5.2. Navedenim algoritmom mogu nastati višestruki poligoni lokalizacija. U tu svrhu, razvijen je odgovarajući algoritam uklanjanja višestrukih lokalizacijskih odziva (engl. *non-maximum supression algorithm*). Algoritam temeljen na sjedinjenim mapama odziva primjenjuje se prilikom lokalizacije velikih objekata kao što su pješački prijelazi (odjeljak 5.3).

*Pojam prenaučenosti odnosi se na pojavu gdje algoritam učenja lokalizacijskog (ili klasifikacijskog) modela uči šum u podacima te generira pretjerano složenu hipotezu. Dobivena hipoteza besprijekorno odvaja uzorke na skupu za učenje, no na skupu za testiranje rezultira znatno lošijom lokalizacijskom performansom. Odjeljak 3.3.1 detaljno opisuje koncept prenaučenosti i standardne metode kojima se pojednostavljuju hipoteze i smanjuje prenaučenost.

U svrhu poboljšanja lokalizacijske preciznosti opisanog sustava slabo nadzirane lokalizacije, dodatno se razmatraju:

- primjena nelinearnih normalizacija Fisherovih vektora slika
- mogućnosti reprezentacije prostornog rasporeda slikovnih riječi.

Nelinearne normalizacije Fisherovih vektora slika Nelinearne normalizacije uključuju normalizaciju potenciranjem (engl. *power normalization*) (u praksi, korijenovanje uz očuvanje predznaka) i metričku normalizaciju [32]. Primjenjuju se na Fisherov vektor cjelokupne slike na temelju kojeg se uči lokalizacijski model. U okviru ove disertacije, uz navedene normalizacije, predlaže se primjena metričke normalizacije po komponentama [45], gdje se doprinos svake slikovne riječi zasebno normalizira. Time se smanjuje „efekt eksplozije slikovne riječi" (engl. *burstiness effect*) [46], gdje pojedina slikovna riječ može imati natprosječno velik doprinos Fisherovu vektoru slike i shodno tome dominirati operacijom skalarnog produkta u odnosu na lokalizacijski model.

Navedene nelinearne normalizacije Fisherova vektora slike invalidiraju aditivnost Fisherovih vektora slikovnih okana, što rezultira činjenicom da se odziv pojedinog okna više ne može izračunati jednostavnim skalarnim produktom u odnosu na lokalizacijski model. U takvom okruženju, izravan proračun doprinosa pojedinog okna uključuje primjenu skupih operacija korjenovanja i dijeljenja za svako okno u slici, što usporava postupak lokalizacije. Kako bi se postigla vremenska učinkovitost izvođenja, predlaže se aproksimacija prvog reda koja odgovara skalarnom produktu nenormaliziranog Fisherova vektora slikovnog okna i gradijenta odziva normalizirane slike u odnosu na slikovno okno. Eksperimenti pokazuju da opisana aproksimacija ne utječe značajno na pad lokalizacijske točnosti, a istovremeno značajno ubrzava postupak lokalizacije.

Reprezentacija prostornog rasporeda Fisherovi vektori podrazumijevaju model slike kao neuređene zbirke slikovnih okana, ne uzimajući u obzir pri tome njihove međusobne odnose. S druge strane, pokazuje se da objekti određenih razreda poprimaju karakterističan globalni ili lokalni prostorni raspored [47]. U okviru područja istraživanja ove disertacije javlja se potreba za modeliranjem lokalnog prostornog rasporeda. Eksperimentalno vrednovanje u okviru odjeljka 5.2.3 pokazuje da opisani sustav lokalizacije uspješno identificira slikovna okna prisutna na objektima traženog razreda. Ponegdje se u okviru pozadinskih objekata sličnih uzoraka kao i u razredu traženog objekta mogu pojaviti okna pozitivnog odziva. Kako bi se uklonili lažni pozitivi i povećavala točnost lokalizacije, predlažu se dva modela reprezentacije lokalnog prostornog rasporeda: prostorni histogrami i prostorni Fisherovi vektori. Navedenim se opisnicima predstavljaju odnosi parova slikovnih riječi za lokalno okruženje svakog pozitivnog okna. Na temelju prostornih opisnika uči se rijedak lokalizacijski model druge razine koji se primjenjuje

za dodatno filtriranje okana koje je model prve razine označio kao pozitivne. Lokalizacijski se poligoni formiraju na temelju svih okana pozitivnog odziva čime se smanjuje broj parametara algoritma generiranja lokalizacijskih poligona.

1.2 Moguće primjene

Područje primjene lokalizacijskih modela učenih uz slabi nadzor uistinu je široko. Velik broj recentnih radova na području slabo nadzirane lokalizacije objekata [18, 48, 49, 50] razmatra taj problem na isključivo konceptualnoj razini bez stvarne primjene na skupovima podataka za koje nisu dostupne lokacije opisanih poligona objekata u primjerima za učenje. Vrednovanje postupaka slabo nadzirane lokalizacije u okviru ove disertacije se, stoga, značajno razlikuje u odnosu na srodne radove na tom području. Razmatraju se primjene lokalizacijskih modela učenih postupcima slabo nadziranog učenja na problemima iz stvarnog života, odnosno na problemima:

- automatizacije sigurnosnih inspekcija prometnica
- automatizacije digitalnog kartiranja.

1.2.1 Automatizacija verifikacije cestovne i komunalne infrastrukture

Kako bi se zadovoljili propisani sigurnosni standardi, poduzeća za upravljanje i održavanje cestovne i komunalne infrastrukture dužna su periodički obavljati provjere. Sigurnosne provjere [51, 52] najčešće se provode u nadležnosti tijela lokalne uprave koja u nedostatku resursa izdaju taj posao (engl. *outsourcing*) vanjskim izvođačima [53]. Tim stručnjaka iz službe ophodnje izlazi na teren, obavlja procjene stanja infrastrukture te u konačnici izdaje službeno izvješće nadležnim institucijama.

U svrhu automatizacije procesa, poželjno je izgraditi geoinformacijski sustav [53, 54, 55, 56, 57] za prikaz elemenata komunalne i cestovne infrastrukture i pametno upravljanje terenskim radovima i intervencijama. Izgradnja sustava za upravljanje prostornim podacima uključuje sljedeće korake [53]:

1. izgradnja geoprostorne baze podataka za bilježenje referentnog stanja lokalne infrastrukture i
2. utvrđivanje ispravnosti usporedbom trenutnog stanja u odnosu na referentno stanje u bazi.

Izgradnja geoprostorne baze podataka uključuje prikupljanje i obradu podataka primjenom vozila opremeljenog kamerom visoke rezolucije i pozicijskim sezorima (engl. *Global Positioning System, GPS*) kojima se prikupljaju globalne koordinate lokacije na kojoj je slika pribavljena. Rezultat prikupljanja podataka je georeferencirani video zapis. Kako bi se izgradila geoprostorna baza, tim eksperata najčešće obrađuje georeferencirani video te određuje elemente lokalne infrastrukture prisutne na određenoj lokaciji. Navedeni proces je vremenski zahtjevan

i podložan greškama zbog toga što i) georeferencirani video može pokrivati površinu od nekoliko stotina ili tisuća kilometara prometnica te ii) postoji velik broj potencijalno relevantnih objekata. Neki od primjera relevantnih objekata uključuju prometne znakove, oznake na kolniku, rasvjetu, stanice javnog prijevoza, klupe u parku, kante za smeće te povijesne spomenike. Poželjno bi, stoga, bilo automatizirati proces metodama računalnog vida.

Cilj utvrđivanja ispravnosti lokalne infrastrukture jest pronaći anomalije poput oštećenih, vegetacijom zaklonjenih ili ukradenih prometnih znakova, nevažećih znakova ograničenja brzine, izbrisanih ili pogrešno iscrtanih oznaka na kolniku, razbijenih tijela javne rasvjete, oštećenih klupa ili kanti za odlaganje otpada. Kao i u koraku izgradnje geoprostorne baze podataka, za proces utvrđivanja ispravnosti također se pribavlja georeferncirani video čijom se obradom i usporedbom u odnosu na referentno stanje utvrđuju anomalije. Vrijeme uloženo od strane stručnjaka da bi se obavila iscrpna i kvalitetna verifikacija i dalje je prepreka u ostvarivanju češćih sigurnosnih pregleda u zadanim financijskim okvirima. Automatizacija usporedbe georeferenciranog video zapisa u odnosu na referentno stanje primjenom računalnog vida smanjila bi troškove i doprinijela učestalosti verifikacije lokalne infrastrukture.

Metoda lokalizacije „odozdo prema gore" primjenom Fisherovih vektora pogodna je za izgradnju geoprostorne baze podataka i za proces utvrđivanja ispravnosti u odnosu na referentno stanje. Svojstva Fisherovih vektora navedena u odjeljku 1.1 i rijetki modeli omogućuju da se u slici identificiraju objekti različitih veličina kao što su razmjerno maleni prometni znakovi (u prosjeku zauzimaju manje od 1 posto površine slike) ili veliki objekti cestovne signalizacije kao što su pješački prijelazi.

1.2.2 Automatizacija digitalne kartografije pomoću masovno prikupljenih geo-podataka

Digitalna kartografija [58, 59] je multidisciplinarno područje koje obuhvaća primjenu računalne tehnologije za prikupljanje, obradu, pohranu i vizualizaciju prostornih podataka. Pojavom servisa poput Google Maps [60] i Google Earth [61] započela je popularizacija digitalne kartografije u široj javnosti, dok je pojavom OpenStreetMap [62] projekta nastala danas najpopularnija slobodna karta svijeta. OpenStreetMap karta se zasniva na dobrovoljno prikupljenim geo-podacima te na satelitskim snimkama visoke rezolucije koje omogućuje servis Bing [63]. Broj registriranih korisnika stalno raste te u trenutku pisanja ovog rada iznosi 3,2 milijuna [64]. Popularizaciji digitalne kartografije doprinijeli su navigacijski uređaji i pametni telefoni opremljeni modulima za globalno pozicioniranje (engl. *Global Positioning System, GPS*). Putem navedenih uređaja korisnici mogu pohraniti dnevnik vožnje u obliku niza GPS koordinata (engl. *GPS track*) te ga potom učitati (engl. *upload*) na OpenStreetMap poslužitelje [65]. Dodatno, danas je većina pametnih telefona opremljena kamerama koje omogućuju pohranu GPS

lokacije na kojoj je fotografija snimljena, dok mobilni programi poput OpenStreetView [66] i Mapillary [67] omogućavaju prikupljanje video zapisa na razini cestovne mreže (engl. *street-view imagery*).

Sustavno prikupljanje i dodavanje specifičnih značajki u digitalnu kartu dugotrajan je proces podložan pogreškama. Primjerice, planovi putovanja (engl. *navigation route*) za teretna i dostavna vozila bitno su različiti u odnosu na planove putovanja osobnih automobila. Postoje brojna ograničenja u obliku prometnih znakova koja propisuju dozvoljene težine, visine, zabrane prijevoza opasnih materijala ili pravna i vremenska ograničenja (primjerice, zabrane prometa teretnim kamionima noćnim satima tijekom turističke sezone). S druge strane, dostavna vozila u pravilu mogu voziti prilaznim cestama do velikih poduzeća i industrijskih zona gdje je na snazi zabrana prometa za osobne automobile. Kako bi se pravilno izračunali planovi putovanja za teretna i dostavna vozila, potrebno je s velikom preciznošću prikupiti podatke o zadanim ograničenjima, što najčešće uključuje prikupljanje podataka na terenu putem GPS uređaja ili putem georeferenciranog video zapisa. Dodatno, sve veći broj gradova diljem svijeta pokreće tzv. projekte „pametnih gradova" (engl. *smart city*) [68], gdje lokalna uprava izdaje programe za pametne telefone ne bi li se bolje prilagodila potrebama građana i podigla kvalitetu života. Primjerice, grad Dubai [69] izdaje besplatne navigacijske programe za pametne telefone koji upozoravaju korisnike o cijenama parkiranja u pojedinim zonama grada te o potencijalnim prometnim gužvama ili zakrčenjima (engl. *Traffic message channel, TMC*). Jasno je da i navigacijski programi za pametne gradove zahtijevaju vrlo precizne kartografske podatke.

Metoda lokalizacije „odozdo prema gore" Fisherovim vektorima primjenljiva za automatizaciju digitalne kartografije. Posebno je zanimljiva sinergija metode lokalizacije „odozdo prema gore" u slabo nadziranom okruženju i masovno prikupljenih geopodataka (engl. *crowdsourcing*) od strane servisa kao što je OpenStreetMap [62] koji se mogu koristiti za generiranje oznaka na razini slike. Za potrebe dobavljanja slikovnog sadržaja mogu se također koristiti dobrovoljno prikupljeni podaci putem servisa OpenStreetView [66] te Mapillary [67]. U okviru ovog rada korišteni su video zapisi prikupljeni u okviru projekata „E-ceste" za Hrvatske gradove Karlovac i Sisak od strane poduzeća „Promet i prostor" [70].

1.3 Znanstveni doprinosi

U okviru ovog poglavlja sažeti su znanstveni doprinosi ostvareni prilikom razvoja slabo nadziranog lokalizacijskog sustava opisanog u odjeljku 1.1. Izvorni znanstveni doprinosi ovog rada su sljedeći:

1. pristup za lokalizaciju objekata zasnovan na reprezentaciji Fisherovim vektorima i slabo nadziranom učenju po komponentama rijetkih lokalizacijskih modela
2. primjena metričke normalizacije po komponentama na Fisherove vektore slika u svrhu

poboljšanja lokalizacijske preciznosti po komponentama rijetkih lokalizacijskih modela
3. učinkovit postupak primjene nelinearnih normalizacija Fisherovog vektora u fazi lokalizacije pomoću gradijenta odziva normalizirane slike
4. reprezentacija prostornog rasporeda slikovnih riječi u vidu prostornih Fisherovih vektora i prostornih histograma.

Prvi doprinos ove disertacije odnosi se na pristup za lokalizaciju objekata zasnovan na reprezentaciji Fisherovim vektorima i slabo nadziranom učenju lokalizacijskih modela rijetkih po komponentama. Uslijed aditivnosti Fisherovih vektora i ekspresivne moći da prilikom usrednjavanja očuvaju i naglase neuobičajene detalje u slici, lokalizacijski se model uči nad Fisherovim vektorima cjelokupnih slika. U fazi testiranja, dobiveni se model primjenjuje za proračun odziva slikovnih okana ne bi li se ustanovila okna odgovorna za prisutnost objekta u slici. Vremenska efikasnost pretraživanja slikovnih okana postiže se primjenom po komponentama rijetkih modela. Na temelju odabranih komponenti modela, iz razmatranja se uklanjaju slikovna okna koja nemaju značajnu vjerojatnost pridruživanja u odnosu na te komponente, a za preostala okna evaluira se tek djelić visokodimenzionalne reprezentacije Fisherovih vektora. Razvijena su dva algoritma prikladna za lokalizaciju velikih i izrazito malenih objekata kojima se na temelju slikovnih okana pozitivnog odziva generiraju prijedlozi lokalizacijskih poligona.

Prednost ovog pristupa u odnosu na srodne radove na području slabo nadzirane lokalizacije [18, 24, 25, 26, 27, 28, 29] jest da ne zahtijeva sužavanje prostora hipoteza primjenom metoda segmentacije [44] ili *objektnosti* [26] kojima se mogu propustiti potencijalne lokacije objekata. Velik broj radova na području slabo nadziranje lokalizacije [18, 22, 24, 25] zasniva se na konceptu učenja zbirki primjeraka (engl. *Multiple Instance Learning, MIL*) [71]. U okviru učenja zbirki primjeraka, slika se predstavlja kao zbirka okana, pri čemu pozitivne slike sadrže barem jedno pozitivno okno (okno koje se nalazi u okviru traženog objekta), a negativne isključivo negativna okna. Proces se odvija iterativno, pri čemu se u svakoj iteraciji obavlja učenje lokalizacijskog modela koji se potom koristi za odabir primjera za učenje za sljedeću iteraciju. Pristup predstavljen u okviru ove disertacije sličan je MIL konceptu na idejnoj razini, budući da se temelji na učenju lokalizacijskog modela na temelju cjelokupnih slika, no nije iterativan.

Drugi i treći doprinos ove disertacije odnose se na primjenu nelinearnih normalizacija Fisherovih vektora u svrhu poboljšanja lokalizacijske učinkovitosti. Poznato je da normalizacije imaju povoljan učinak prilikom i) određivanja prisutnosti objekata u slikama [32, 33] te ii) lokalizacije objekata u paradigmi strogo nadziranog učenja [6]. U okviru ove disertacije pokazano je da se nelinearne normalizacije mogu uspješno integrirati u razvijeni sustav lokalizacije. Kao drugi doprinos ove disertacije predložena je primjena metričke normalizacije po komponentama na Fisherove vektore slika u fazi učenja po komponentama rijetkih lokalizacijskih modela. Navedena normalizacija uzima u obzir strukturu Fisherova vektora te poništava potencijalni negativni „efekt eksplozije slikovne riječi" [46]. U tom slučaju određena slikovna riječ

daje natprosječno velik doprinos Fisherovu vektoru slike i shodno tome dominira nad operacijom skalarnog produkta u odnosu na klasifikacijski model.

Budući da nelinearne normalizacije invalidiraju aditivnost Fisherovih vektora slikovnih okana, kao treći doprinos ove disertacije predložena je aproksimacija prvog reda za izračun odziva slikovnog okna. Odziv okna tada se određuje na temelju skalarnog produkta gradijenta odziva normalizirane slike i nenormaliziranog Fisherova vektora okna. Na taj se način nelinearne normalizacije primjenjuju samo prilikom proračuna gradijenta, što značajno ubrzava proces lokalizacije.

Četvrti doprinos uključuje izgradnju modela lokalnog prostornog rasporeda. Predložena su dva opisnika kojima se modeliraju odnosi parova slikovnih riječi: prostorni histogrami i prostorni Fisherovi vektori. Prostorni se opisnici formiraju kao opisnici druge razine na temelju okana pozitivnog odziva dobivenih primjenom lokalizacijskog modela učenog nad cjelokupnim slikama. Lokalizacijski se poligoni formiraju na temelju svih okana pozitivnog odziva odabranih od strane lokalizacijskih modela druge razine čime se smanjuje broj parametara algoritma generiranja lokalizacijskih poligona (u slučaju modela prve razine, koristi se top T okana).

Provedeno je iscrpno eksperimentalno vrednovanje opisanih modela lokalizacije i reprezentacije na područjima primjene opisanim u okviru odjeljka 1.2. U okviru automatizacije sigurnosnih inspekcija prometnica, vrednovana je lokalizacija prometnih znakova na javno dostupnom skupu podataka [72]. Prometni su znakovi odabrani kao reprezentativan primjer cestovne infrastrukture te kao primjer malenog objekta koji u prosjeku zauzima manje od jedan posto površine slike. Najmanji primjeri prometnih znakova u okviru korištenog skupa podataka [72] zauzimaju čak manje od dva promila površine slike. Lokalizacija tako malenih objekata predstavlja glavni izazov za učenje lokalizacijskog modela pod slabim nadzorom. U okviru automatizacije digitalne kartografije poluautomatski je prikupljen nov skup podataka na temelju geografskih lokacija pješačkih prijelaza (zemljopisne širine i dužine) iz OpenStreetMap karte [62] i georeferenciranih video zapisa [70]. Razvijen je algoritam automatskog uparivanja GPS (engl. *Global positioning system*) lokacija objekata iz OpenStreetMap karte u odnosu na okvire (engl. *frames*) georeferenciranog video zapisa. S obzirom da OpenStreetMap karta sadrži podatke prikupljene radom mnoštva (engl. *crowdsourcing*) i GPS uređaja različite preciznosti, primjenom razvijenog algoritma prikupljene su i slike koje ne sadrže pješačke prijelaze. Analizom je ustanovljeno da 15 posto prikupljenih slika ne sadrži pješačke prijelaze koje su potom uklonjene ručnim probirom. Dobiveni skup podataka odlikuje vrlo velika unutar-razredna varijabilnost, gdje su pješački prijelazi snimljeni iz različitih kutova gledanja (sprijeda, bočno), različitih udaljenosti u odnosu na kameru te različitih osvjetljenja (uslijed vremenskih uvjeta). Dodatno, oko 80 posto video materijala prikupljeno je u urbanim sredinama, gdje se uz pješačke prijelaze pojavljuje mnoštvo sličnih pozadinskih objekata kao što su pješački otoci i drugi oblici cestovne signalizacije koji dodatno otežavaju proces učenja lokalizacijskog modela pod slabim

nadzorom. Zbog opisanih problema, ocijenjeno je da je zadatak lokalizacije pješačkih prijelaza težak i zanimljiv problem. Dobiveni eksperimentalni rezultati lokalizacije pješačkih prijelaza imaju, stoga, konceptualni značaj: pokazuju da se metode računalnog vida uspješno mogu koristiti za obogaćivanje javno dostupnih kartografskih podataka novim objektima.

Opisani pristup lokalizaciji vrednovan je za dva tipa opisnika slikovnih okana, konkretno za SIFT značajke [37] i konvolucijske značajke [39]. Rezultati pokazuju da je pristup prikladan za lokalizaciju izrazito malenih objekata kao što su prometni znakovi, ali i velikih objekata sa velikom unutar razrednom varijacijom kao što su pješački prijelazi. Vrednovanjem navedenih doprinosa ustanovljeno je:

- Modeli rijetki po komponentama postižu bolju vremensku učinkovitost i lokalizacijsku preciznost u odnosu na guste ℓ_2 modele i rijetke ℓ_1 modele.
- Metrička normalizacija po komponentama u konjunkciji sa po komponentama rijetkim modelima poboljšava efikasnost izvođenja i lokalizacijsku točnost.
- Uslijed primjene nelinearnih normalizacija povećava se stupanj rijetkosti lokalizacijskog modela te se primjenom predložene aproksimacije prvog reda postiže bolja vremenska učinkovitost čak i u odnosu na slučaj bez normalizacija.
- Aproksimacija gradijentom ne utječe značajno na pad lokalizacijske točnosti.
- Lokalizacijskim modelom druge razine učenim na temelju prostornih opisnika uklanjaju se lažni pozitivi i poboljšava lokalizacijska točnost.

1.4 Struktura rada

U okviru ovog poglavlja opisana je struktura ove disertacije.

Drugo poglavlje pruža uvid u do sada postignute rezultate u srodnim istraživačkim područjima. Poglavlje je podijeljeno u nekoliko cjelina. U prvoj se cjelini razmatraju postojeće vektorske reprezentacije slikovnih okana i slika: i) opisnici slikovnih okana (tzv. lokalni opisnici) te ii) opisnici iz porodice zbirke slikovnih riječi. U drugoj se cjelini razmatraju mogućnosti predstavljanja prostornog rasporeda za opisnike temeljene na zbirkama slikovnih riječi. U trećoj cjelini se razmatraju srodni radovi na području lokalizacije objekata. Detaljno se opisuju postupci pretraživanja i sužavanja prostora mogućih lokacija objekata, odnosno različite strategije lokalizacije. Budući da je slabo nadzirana lokalizacija jedan od glavnih doprinosa ove disertacije, u okviru četvrte cjeline detaljno se razmatraju srodni radovi na području slabo nadziranog učenja u računalnom vidu.

U trećem poglavlju formalizira se koncept „lokalizacije odozdo prema gore" i detaljno se opisuju sve komponente sustava lokalizacije. Poglavlje uključuje detaljan opis kôdiranja Fisherovim vektorima, normalizacija Fisherovih vektora, diskriminativnih linearnih modela i regularizacijskih funkcija koje induciraju rjetkoću modela. Također je predstavljena aproksimacija

odziva okna pomoću gradijenta odziva normalizirane slike. Navedena aproksimacija omogućuje učinkovito računanje odziva slikovnog okna uz primjenu nelinearnih normalizacija u fazi učenja. Predstavljeni su postupci kojima se ubrzava pretraživanja slikovnih okana u fazi lokalizacije, algoritmi formiranja lokalizacijskih poligona na temelju slikovnih okana i algoritmi za uklanjanje višestrukih poligona lokalizacija.

Četvrto poglavlje opisuje postupak izgradnje modela reprezentacije lokalnih prostornih odnosa među slikovnim riječima. Predstavljena su dva opisnika lokalnog rasporeda: prostorni histogrami i prostorni Fisherovi vektori. Opisane su optimizacije izračuna prostornih Fisherovih vektora.

Peto poglavlje uključuje detaljno eksperimentalno vrednovanje predloženih doprinosa disertacije na dva problema iz stvarnog svijeta: automatizaciji pregleda cestovne infrastrukture i automatizaciji digitalne kartografije. U okviru automatizacije pregleda cestovne infrastrukture razmatra se lokalizacija prometnih znakova, a u okviru automatizacije digitalne kartografije, lokalizacija pješačkih prijelaza. Opisan je algoritam stvaranja novog skupa slika pješačkih prijelaza označenih informacijom o prisutnosti objekta u slici. Algoritam se temelji na uparivanju masovno prikupljenih lokacija pješačkih prijelaza iz OpenStreetMap karte u odnosu na georeferencirane video zapise.

Konačno, posljednje šesto poglavlje zaključuje distertaciju dajući sažet pregled predstavljenog istraživanja. U okviru ovog poglavlja analizira se značaj pojedinih doprinosa te se razmatraju moguća poboljšanja i daljnji rad.

Poglavlje 2

Pregled srodnih istraživačkih područja

S obzirom na znanstvene doprinose navedene u okviru odjeljka 1.3, pregled srodnih istraživačkih područja podijeljen je u nekoliko cjelina. U odjeljku 2.1 razmatraju se srodni radovi na području reprezentacije slika. Posebna je pažnja usmjerena na reprezentacije iz porodice zbirke slikovih riječi kojoj pripadaju Fisherovi vektori. U odjeljku 2.2, dan je pregled recentnih radova koji razmatraju problem predstavljanja prostornog rasporeda. Različiti pristupi podijeljeni su u skupine s obzirom na opseg prostornog rasporeda koji opisuju. U okviru odjeljka 2.3 razmatraju se različite strategije lokalizacije te njihov odnos s obzirom na lokalizaciju „odozdo prema gore" razmatranu u okviru ove disertacije. Kako je u okviru ove disertacije posebna pažnja usmjerena na problem slabo nadzirane lokalizacije, u odjeljku 2.4 dan je pregled značajnih radova na tom području.

2.1 Prikaz slike vektorom značajki

U prethodnom su poglavlju opisani neki od temeljnih zadataka računalnog vida: klasifikacija slika i lokalizacija objekata u slikama. Prilikom rješavanja tih problema, pogodno je sliku prikazati vektorom značajki [73]. Primjerice, tada se za klasifikaciju i lokalizaciju mogu koristiti algoritmi strojnog učenja kao što su logistička regresija ili stroj s potpornim vektorima [34, 42]. Kako se u okviru ove disertacije koristi se reprezentacija Fisherovim vektorima [31, 33] posebna je pažnja usmjerena na opisnike porodice zbirke slikovnih riječi (engl. *Bag of Visual Words, BoVW*) [40, 41] kojoj Fisherovi vektori pripadaju. Motivacija za reprezentacijom slike zbirkom slikovnih riječi u računalnom vidu prvi puta se pojavila prilikom razmatranja problema dohvata objekata iz video zapisa (engl. *object retrieval*) [41]. Sivic i Zisserman u okviru rada [41] pristupaju problemu dohvata objekata iz video zapisa razmatrajući metode koje se koriste za analizu i pretraživanje teksta. Sasvim općenito, za potrebe analize i pretraživanja teksta (engl. *text retrieval*), pojedini dokument se parsira i za svaku riječ se određuje korijen (engl. *stem*). Zatim se uklanjaju česte riječi (primjerice, članovi poput „the" ili „an") koje ne doprinose diskrimi-

nativnosti dokumenta. Preostalim riječima dodjeljuje se jedinstveni identifikator i dokument se predstavlja frekvencijom pojavljivanja pojedinih riječi, odnosno histogramom. Mogućnosti primjene opisanog koncepta iz analize teksta za potrebe klasifikacije slika prvi puta su razmatrane u okviru rada [40], a od tada se odgovarajuća proširenja i poboljšanja primjenjuju iznimno često za zadatke pretraživanja, klasifikacije slika i lokalizacije objekata [31, 74, 75, 76].

U nastavku odjeljka opisana su svojstva i koraci izgradnje opisnika porodice zbirke slikovnih riječi. Navedeni su recentni radovi i poboljšanja svih faza izgradnje opisnika zbirke slikovnih riječi s naglaskom na usporedbu s reprezentacijom Fisherovih vektora koji predstavljaju jedan od ključnih elemenata ove disertacije. Sasvim općenito, proces izgradnje opisnika slike zbirkom slikovnih riječi može se podijeliti u četiri koraka ilustrirana u slici 2.1:

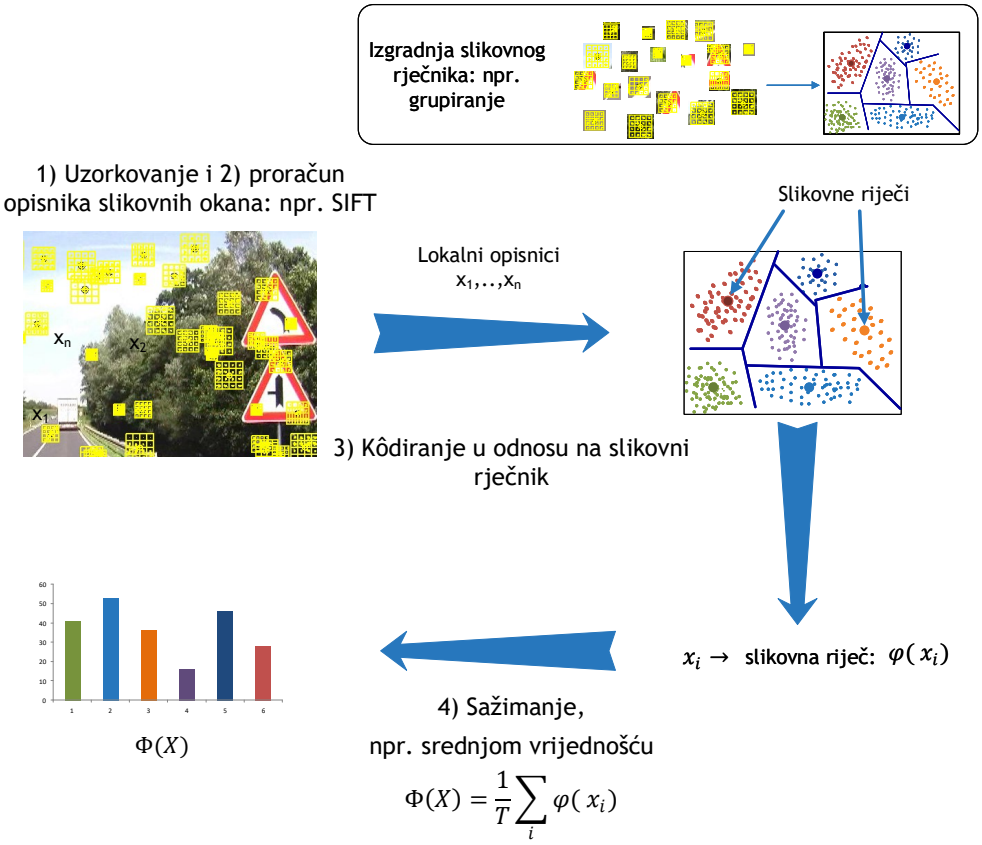
1. uzorkovanje slikovnih okana,
2. izračun opisnika niske razine za slikovna okna,
3. kôdiranje opisnika slikovnih okana u odnosu na prethodno naučeni slikovni rječnik te
4. sažimanje kôdiranih lokalnih opisnika u konačan opisnik slike.

2.1.1 Uzorkovanje slikovnih okana

Važna zadaća procesa uzorkovanja slikovnih okana jest dohvatiti reprezentativan skup okana koji predstavlja relevantnu informaciju o slici. Sasvim općenito, u literaturi se mogu pronaći različiti tipovi uzorkovanja slikovnih okana, no najčešće primjenjivani su [77]: i) uzorkovanje interesnih točaka te ii) gusto uzorkovanje na različitim mjerilima.

Uzorkovanje interesnim točkama Cilj uzorkovanja interesnih točaka jest pronaći skup lokacija invarijantnih na promjene perspektive (primjerice skaliranjem, translacijom ili rotacijom) i osvjetljenja. U literaturi se mogu pronaći sljedeći algoritmi uzorkovanja interesnih točaka:

- *Harris-Laplace* algoritam neosjetljiv na promjene u mjerilu i *Harris-Affine* algoritam neosjetljiv na afine transformacije objekta predloženi su u [78] kao proširenje originalnoga Harrisova algoritma za detekciju kuteva [79] koji se temelji na autokorelacijskoj matrici i tzv. mjeri *kutnosti* značajke (engl. *cornerness*);
- Algoritam pronalaženja lokalnih ekstrema u piramidi razlika Gaussovih filtara (engl. *Difference of Gaussian*) [37] pronalazi interesne točke neosjetljive na promjene u mjerilu.
- Algoritam pronalaženja regija najveće stabilnosti [80] (engl. *Maximally Stable Extremal Regions, MSER*) pronalazi povezane komponente nad sivom slikom (engl. *grayscale*) na koju je primijenjen određeni prag (engl. *threshold*). Komponente koje bilježe najmanju promjenu površine s obzirom na variranje vrijednosti piksela smatraju se regijama najveće stabilnosti. Dobivene regije neosjetljive su na afine geometrijske i fotometrijske transformacije.



Slika 2.1: Shematski prikaz izgradnje histograma slikovnih riječi (engl. *Bag of visual words, BoVW*).

Gusto uzorkovanje Algoritmi gustog uzorkovanja obrađuju sve piksele u slici te izdvajaju razmjerno malena i međusobno preklapajuća okna na različitim mjerilima (engl. *multi-scale dense grid*). Autori u [77] su pokazali da gusto uzorkovanje daje bolje rezultate za zadatke klasifikacije slika i lokalizacije objekata, dok je uzorkovanje interesnih točaka efikasno za prepoznavanje istog fizičkog objekta u različitim slikama (engl. *object instance recognition*). Objekti se u slikama pojavljuju na različitim mjestima te u različitim veličinama (mjerilima) i stoga okna ključna za klasifikaciju, odnosno lokalizaciju nisu nužno povezana sa interesnim točkama. Problem gustog uzorkovanja je u računalnoj složenosti samog procesa. Naime, kako bi se identificirala slična okna u različitim slikama, potrebno je primijeniti razmjerno veliku frekvenciju uzorkovanja. Iz tog je razloga iznimno važno da proračun opisnika za okna dobivena gustim uzorkovanjem bude računski efikasan.

2.1.2 Opisnici niske razine

U drugom koraku postupka za izvođenje opisnika zbirke slikovnih riječi, prikazanog na slici 2.1, obavlja se izračun opisnika slikovih okana dobivenih gustim uzorkovanjem ili uzorkovanjem interesnih točaka. Opisnici slikovnih okana nazivaju se još i opisnicima niske razine ili

lokalnim opisnicima te predstavljaju slikovno okno vektorom koji je neosjetljiv u odnosu na lokalne fotometrijske i geometrijske promjene. Sasvim općenito, opisnike niže razine možemo podijeliti na opisnike ugađane za posebnu namjenu (engl. *hand-crafted descriptors*) i opisnike dobivene učenjem (engl. *learned local descriptors*). Najvažniji predstavnici opisnika ugađanih za posebnu namjenu uključuju Scale Invariant Feature Transform (SIFT) [37], Speeded Up Robust Features (SURF) [81], opisnik lokalnih binarnih uzoraka (engl. *Local Binary Patterns, LPB*) [82] te histogram orijentiranih gradijenata (engl. *Histogram of Oriented Gradients, HOG*) [4]. Konvolucijske značajke [18, 83, 84, 85, 86, 87] predstavljaju primjer opisnika dobivenih učenjem.

Scale Invariant Feature Transform (SIFT)

Opisnik SIFT [37, 102] jedan je od najčešće korištenih opisnika niske razine. Primjenjuje se u kombinaciji s gustim uzorkovanjem [6, 20, 88] i uzorkovanjem interesnih točaka [37, 89]. U okviru eksperimentalnog vrednovanja lokalizacije prometnih znakova u odjeljcima 5.2.2 i 5.2.3 koriste se SIFT značajke [88, 102] fiksne orijentacije gusto uzorkovane na više mjerila (engl. *dense SIFT*). Izgradnja samog opisnika uzorkovanog na određenom mjerilu obavlja se sljedećim koracima:

1. Za svaki piksel u okviru okna izračuna se gradijent određen amplitudom i smjerom.
2. Okno se dijeli u pravokutnu mrežu od 4×4 ćelije.
3. Na razini svake ćelije, gradijenti pojedinih piksela pohranjuju se u histogram od 8 odjeljaka koji predstavljaju usmjerenja u intervalu $[0, 180°]$, pri čemu je doprinos pojedinog piksela odgovarajućem odjeljku histograma određen amplitudom gradijenta.
4. Histogrami pojedinih ćelija nadovezuju se u konačni vektor dimenzije 128 elemenata ($4\times4\times8$).
5. Dobiveni opisnik se ℓ_2 normalizira.
6. Postavlja se prag na amplitude gradijenata najčešće na iznos 0.2 nakon čega se opisnik ponovno ℓ_2 normalizira.

Opisnik SIFT postiže neosjetljivost na različite promjene opisane u nastavku. Korištenjem gradijenata umjesto intenziteta postiže se neosjetljivost na lokalnu razinu svjetline. Primjenom ℓ_2 normalizacije u petom koraku izgradnje opisnika, postiže se neosjetljivost na promjene u kontrastu i linearne promjene u osvjetljenju. Nelinearne promjene poput zasićenja te nelinearne promjene u osvjetljenju pridonose velikim promjenama u amplitudi gradijenta. Kako bi se postigla neosjetljivost na potencijalni utjecaj takvih promjena, u koraku šest izgradnje opisnika postavlja se prag od 0.2 na iznos amplitude gradijenta.

Konvolucijske značajke

Konvolucijske neuronske mreže trenutno postižu najbolje rezultate na području klasifikacije i segmentacije slika te lokalizacije objekata [7, 90, 91, 92]. Osim za probleme klasifikacije i lokalizacije, izlazi pojedinih slojeva konvolucijske mreže također se upotrebljavaju u domeni prijenosa učenja (engl. *transfer learning*). Koncept prijenosa učenja nema duboke teorijske temelje, no u praksi daje iznimno dobre rezultate [18, 84, 85, 86, 87]. Osnovna ideja je učiti konvolucijsku mrežu minimizacijom klasifikacijskog gubitka na velikom skupu podataka kao što je primjerice ImageNet ILSVRC [93]. Izlaz odabranog sloja mreže tada se može koristiti za reprezentaciju slike [83]. U prilog činjenici da se izlazi međuslojeva mreže mogu koristiti za reprezentaciju slike, govori i rad predstavljen u [90], gdje autori, između ostalog, razmatraju problem vizualizacije izlaza pojedinih slojeva mreže. Pokazuje se da odzivi konvolucijskih slojeva mreže (mape značajki) odgovaraju strukturama više razine u slici, kao što su primjerice lica ili tijela životinja.

Sasvim općenito, konvolucijsku mrežu možemo prikazati kao kompoziciju M slojeva $\phi_M \circ \cdots \phi_2 \circ \phi_1$, gdje M odgovara dubini mreže. Određeni sloj ϕ_m uključuje C_m neurona (kanala), od kojih svaki na izlazu daje mapu značajki dimenzija $W_m \times H_m$. S porastom dubine mreže, smanjuje se prostorna rezolucija mapa značajki $W_m \times H_m$, a povećava se broj kanala u slojevima C_m. Tipovi slojeva u konvolucijskoj mreži uključuju:

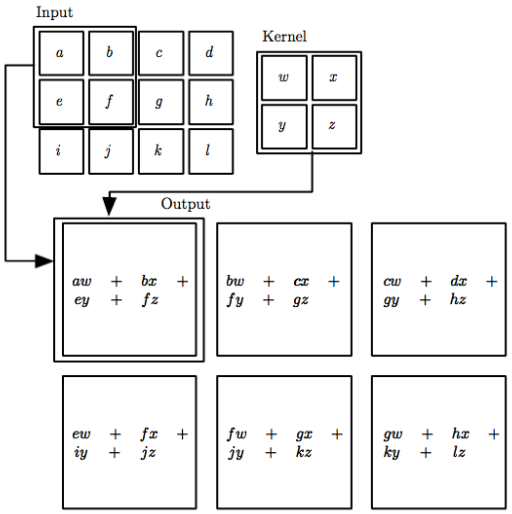
- konvolucijski sloj (engl. *convolutional layer*), nakon kojeg obično slijedi nelinearna aktivacijska funkcija
- sloj sažimanja (engl. *pooling layer*)
- potpuno povezani sloj (engl. *fully connected layer*).

Konvolucijski sloj U okviru konvolucijskog sloja (engl. *convolutional layer*) provodi se operacija konvolucije ulazne mape značajki I s jezgrom K dimenzija $K_w \times K_h$. Na samom ulazu mreže, ulazna mapa značajki I odgovara matrici s intenzitetima piksela slike. Operacija konvolucije sa jezgrom K obavlja se u maniri posmičnog prozora (slika 2.2), gdje su matrice jezgre najčešće kvadratne. Pomak jezgre definiran je parametrima S_w po širini i S_h po visini. Sam rezultat konvolucije na poziciji (i, j) definiran je kao:

$$(I * K)(i, j) = \sum_{w=w_{min}}^{w_{max}} \sum_{h=h_{min}}^{h_{max}} I(i+w, j+h) K(w, h) \, . \tag{2.1}$$

Za granice gornjih operatora sumacije vrijedi $K_w = w_{max} - w_{min}$, te $K_h = h_{max} - h_{min}$. Kao rezultat prolaza jezgre K preko ulaza I dimenzija $W_{m-1} \times H_{m-1}$ u sloju ϕ_{m-1}, dobivaju se mape značajki dimenzija $W_m \times H_m$, pri čemu:

$$W_m = \frac{W_{m-1} - K_w}{S_w} + 1, \; H_m = \frac{H_{m-1} - K_h}{S_h} + 1 \, . \tag{2.2}$$



Slika 2.2: Primjer prolaza jezgre K, dimenzija $K_w = K_h = 2$ sa pomakom $S_w = S_h = 1$. Preuzeto iz [94].

Operacija konvolucije ekvivarijantna je u odnosu na pomak, odnosno ako pomaknemo ulaz I, pomiče se i mapa značajki.

Na izlazu konvolucijskog sloja primjenjuje se aktivacijska (prijenosna) funkcija. Najčešći tipovi aktivacijskih funkcija uključuju:

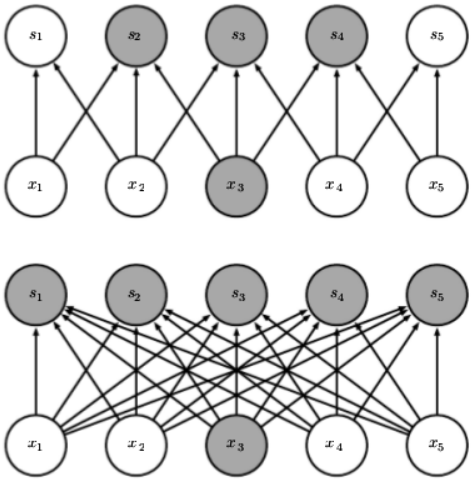
- funkciju zgloba (engl. *rectified linear unit*) $\text{ReLU}(x) = \max(0, x)$
- sigmoidnu funkciju $1/[1 + \exp(-x)]$
- tangens hiperbolni $[exp(2x) - 1]/[exp(2x) + 1]$.

Uloga aktivacijske funkcije je unošenje nelinearnosti u mrežu.

Sloj sažimanja Sloj sažimanja (podotipkavajući sloj) (engl. *pooling layer*) smanjuje rezoluciju mapa značajki, odnosno preslikava skup prostorno bliskih značajki na ulazu u jednu značajku na izlazu. Prozorom veličine $P_w \times P_h$ prolazi se preko ulazne mape i vrijednosti unutar prozora zamjenjuju se jednom vrijednošću koja na izlazu predstavlja čitav prozor. Uloga sloja sažimanja jest povećati neosjetljivost na pomak jer se smatra da je zadatak određivanja prisutnosti objekta važniji u odnosu na njegovu točnu poziciju. U odnosu na samu poziciju značajke, važniji je međusobni prostorni raspored značajki. Neki od primjera funkcija sažimanja uključuju:

- sažimanje maksimalnom vrijednošću (engl. *max pooling*)
- sažimanje usrednjavanjem (engl. *average pooling*)
- sažimanje ℓ_2 normom
- sažimanje usrednjavanjem pomoću Gaussove jezgre.

Potpuno povezani sloj Potpuno povezani sloj (engl. *fully connected layer*) povezuje sve ulaze na sve izlaze sloja. Slika 2.3 ilustrira razlike između konvolucijskog i potpuno povezanoga sloja.



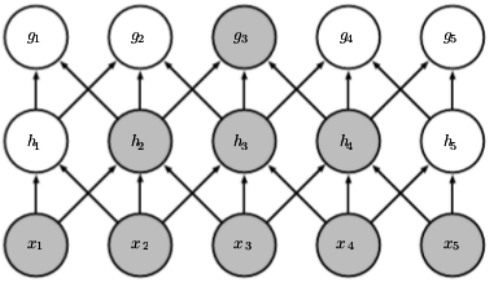
Slika 2.3: Razlika konvolucijskog sloja (gornji redak) u odnosu na potpuno povezani sloj (donji redak). Preuzeto iz [94].

U odnosu na konvolucijski sloj, potpuno povezani sloj ima znatno veći broj parametara (težina) koje je potrebno naučiti te ga shodno tome karakterizira i sporija evaluacija. Konkretno, neka je u jednodimenzionalnom slučaju, k širina jezgre konvolucijskog sloja, a broj neurona (kanala) označen je s n, tada je složenost izvođenja konvolucijskog sloja $O(kn)$. S druge strane, neka je m broj ulaza potpuno povezanog sloja, a n broj neurona u potpuno povezanom sloju, složenost izvođenja potpuno povezanog sloja iznosi $O(mn)$. Budući da vrijedi $k \ll m$, složenost izvođenja konvolucijskog sloja znatno je manja $O(kn) \ll O(mn)$.

Naposljetku, valja napomenuti da recentni radovi na području učenja „s kraja na kraj" (engl. *end-to-end learning*) [95, 96] nakon posljednjeg konvolucijskog sloja predlažu globalno sažimanje srednjom ili maksimalnom vrijednošću umjesto niza potpuno povezanih slojeva. Globalnim se sažimanjem svaka mapa značajki predstavlja skalarom koji označava srednju ili maksimalnu vrijednost svih elemenata. Dobiveni vektori značajki dovode se na ulaz potpuno povezanog sloja čiji broj neurona odgovara broju razreda klasifikacije koja se mrežom provodi.

Receptivno polje Receptivno polje određenog sloja mreže ϕ_m obuhvaća značajke na ulazu mreže koje neposredno mogu utjecati na vrijednosti neurona u sloju ϕ_m. Na taj način dublji slojevi mreže mogu predstavljati veće površine ulazne slike. Slika 2.4 ilustrira primjer receptivnog polja neurona u drugom skrivenom sloju neuronske mreže.

Arhitektura duboke konvolucijske mreže VGG-E U okviru ovog rada koriste se konvolucijske značajke izvedene iz VGG-E mreže [39]. Mreža VGG-E broji ukupno 19 slojeva i shematski je prikazana krajnje desno na slici 2.5. Ulazna slika najprije prolazi kroza stog konvolucijskih slojeva u okviru kojih se primjenjuje jezgra rezolucije $K_w = K_h = 3$ s pomakom od $S_w = S_h = 1$ piksela. Na izlazu svakog konvolucijskog sloja obavlja se aktivacijska funkcija



Slika 2.4: Primjer receptivnog polja drugog sloja mreže: receptivno polje neurona g_3 čine ulazi mreže x_1, x_2, x_3, x_4 i x_5. Preuzeto iz [94].

tipa zglobnice (ReLU) kako bi se postigla nelinearnost. Mreža uključuje ukupno pet slojeva sažimanja u kojima se obavlja sažimanje maksimalnom vrijednošću (engl. *max pooling*). Sažimanje se obavlja nad prozorom dimenzija $P_w = P_h = 2$ sa pomakom od 2 piksela. Na taj se način dimenzionalnost mape značajki smanjuje dva puta u odnosu na ulaznu mapu značajki.

U okviru eksperimentalnog vrednovanja slabo nadzirane lokalizacije pješačkih prijelaza u odjeljku 5.3 koriste se mape značajki dobivene kao izlaz posljednjeg konvolucijskog sloja conv3-512. Dani sloj broji 512 mapa značajki, a receptivno polje iznosi 252×252 piksela.

2.1.3 Obrada opisnika niske razine

Nakon izdvajanja lokalnih opisnika, a prije samog postupka kôdiranja, može se primijeniti neka od tehnika smanjenja dimenzionalnosti. Jedna od najčešće [20, 33, 97] primjenjivanih metoda jest analiza glavnih komponenata (engl. *Principal components analysis, PCA*) [34, 42]. Postupkom analize glavnih komponenti skup lokalnih opisnika $\mathbf{X} = \{\mathbf{x_t},\ t = 1 \ldots T\}$, gdje dimenzionalnost lokalnih opisnika odgovara D, $\mathbf{x_t} \in \mathbb{R}^D$, projicira se u d-dimenzionalni prostor, gdje $d < D$. Smanjenjem dimenzionalnosti smanjuje se zalihost u podacima, odnosno obavlja se dekorelacija pojedinih dimenzija lokalnih opisnika. Lokalni se opisnici projiciraju u ciljani potprostor niže dimenzionalnosti na način da se maksimizira varijanca projiciranih podataka. Neka $\mathbf{u}$ označava glavnu os varijacije, odnosno jedinični vektor na koji želimo projicirati podatke. Duljina projekcije opisnika $\mathbf{x_t}$ na vektor $\mathbf{u}$ dana je sa $\mathbf{x_t}^T\mathbf{u}$. Kako bismo maksimizirali varijancu podataka, potrebno je maksimizirati kriterij [98]:

$$J = \frac{1}{T}\sum_{t=1}^{T}(\mathbf{x_t}^T\mathbf{u})^2 = \mathbf{u}^T\left(\frac{1}{T}\sum_{t=1}^{T}\mathbf{x_t}\mathbf{x_t}^T\right)\mathbf{u}\,. \tag{2.3}$$

Maksimizacija izraza (2.3) uz ograničenje $\|\mathbf{u}\| = 1$ daje svojstvene vektore empirijske matrice korelacije lokalnih opisnika $\Sigma = \frac{1}{T}\sum_{t=1}^{T}\mathbf{x_t}\mathbf{x_t}^T$. Kako bismo lokalne opisnike projicirali u potprostor dimenzije d, potrebno je odabrati d svojstvenih vektora $\mathbf{u_1} \ldots \mathbf{u_d}$ matrice $\mathbf{\Sigma}$ kojima od-

ConvNet Configuration					
A	A-LRN	B	C	D	E
11 weight layers	11 weight layers	13 weight layers	16 weight layers	16 weight layers	19 weight layers
input (224 × 224 RGB image)					
conv3-64	conv3-64 **LRN**	conv3-64 **conv3-64**	conv3-64 conv3-64	conv3-64 conv3-64	conv3-64 conv3-64
maxpool					
conv3-128	conv3-128	conv3-128 **conv3-128**	conv3-128 conv3-128	conv3-128 conv3-128	conv3-128 conv3-128
maxpool					
conv3-256 conv3-256	conv3-256 conv3-256	conv3-256 conv3-256	conv3-256 conv3-256 **conv1-256**	conv3-256 conv3-256 **conv3-256**	conv3-256 conv3-256 conv3-256 **conv3-256**
maxpool					
conv3-512 conv3-512	conv3-512 conv3-512	conv3-512 conv3-512	conv3-512 conv3-512 **conv1-512**	conv3-512 conv3-512 **conv3-512**	conv3-512 conv3-512 conv3-512 **conv3-512**
maxpool					
conv3-512 conv3-512	conv3-512 conv3-512	conv3-512 conv3-512	conv3-512 conv3-512 **conv1-512**	conv3-512 conv3-512 **conv3-512**	conv3-512 conv3-512 conv3-512 **conv3-512**
maxpool					
FC-4096					
FC-4096					
FC-1000					
soft-max					

Table 2: **Number of parameters** (in millions).

Network	A,A-LRN	B	C	D	E
Number of parameters	133	133	134	138	144

Slika 2.5: Shematski prikaz arhitektura duboke konvolucijske mreže VGG [39]. Shema mreže VGG-E i odgovarajući broj parametara prikazani su u krajnje desnom stupcu.

govaraju najveće svojstvene vrijednosti. Operacija projekcije definirana je izrazom:

$$\mathbf{y_t} = \begin{bmatrix} \mathbf{u_1}^T \mathbf{x_t} \\ \mathbf{u_2}^T \mathbf{x_t} \\ \vdots \\ \mathbf{u_d}^T \mathbf{x_t} \end{bmatrix}. \tag{2.4}$$

2.1.4 Slikovni rječnik

Slikovni rječnik (engl. *visual vocabulary, codebook*) predstavlja statistiku općenitog sadržaja slike te kao takav predstavlja ključnu komponentu u procesu izgradnje opisnika porodice slikovnih riječi. Proces kôdiranja opisnika slikovnih okana u višedimenzionalni prostor kôdova obavlja se pridruživanjem u odnosu na slikovni rječnik. Postupci pridruživanja potanko su opisani u sljedećem odjeljku. Neki od primjera slikovnih rječnika koji se mogu pronaći u literaturi uključuju:

- skup centroida pojedinih grupa dobivenih grupiranjem prema algortimu *k*-srednjih vrijednosti (engl. *k-means clustering*) [42, 99, 100]
- model mješavine Gaussovih razdiobi (engl. *Gaussian mixture model, GMM*) [42]
- slučajne distribucijske šume (engl. *Random density forests, RDF*) [101].

Za sadržaj ove disertacije relevantni su Fisherovi vektori koji predstavljaju proširenje histograma slikovnih riječi (engl. Bag of Visual Words, BoVW). Iz tog razloga, pripadajući slikovni rječnici detaljnije će biti opisani u nastavku.

Grupiranje algoritmom *k*-srednjih vrijednosti Grupiranje algoritmom k-srednjih vrijednosti (engl. *k-means clustering*) [42, 99, 100] koristi se za izgradnju slikovnog rječnika kod histograma slikovnih riječi BoVW. Algoritam k-srednjih vrijednosti pripada paradigmi učenja bez nadzora. Za skup lokalnih opisnika $\mathbf{X} = \{\mathbf{x_t},\ t = 1 \dots T\}$ i zadanu veličinu slikovnog rječnika K, algoritmom se lokalni opisnici grupiraju na način da udaljenost opisnika unutar grupe bude manja u odnosu na udaljenosti izvan grupe. Svaka grupa opisuje se reprezentom $\boldsymbol{\mu}_k$ (slikovnom riječi), a pripadnost opisnika $\mathbf{x_t}$ grupi k bilježi se binarnom varijablom $q_{tk} \in \{0,1\}$. Varijabla q_{tk} poprima vrijednost jedinice ukoliko je $\mathbf{x_t}$ dodijeljen k-toj grupi, u suprotnom je jednaka ništici. Cilj algoritma učenja je odrediti vrijednosti varijabli q_{tk}, $t = 1 \dots T$, $k = 1 \dots K$, te reprezenata grupi $\boldsymbol{\mu}_k$. To se postiže minimizacijom sljedećeg gubitka [42]:

$$J = \sum_{t=1}^{T} \sum_{k=1}^{K} q_{tk} \|\mathbf{x_t} - \boldsymbol{\mu}_k\|^2 \,. \tag{2.5}$$

Funkcija gubitka (2.5) predstavlja sumu kvadrata udaljenosti svakog opisnika u odnosu na njemu dodijeljenu slikovnu riječ $\boldsymbol{\mu}_k$. Algoritam optimizacije započinje inicijalizacijom gdje se određuju početne vrijednosti slikovnih riječi $\boldsymbol{\mu}_k$, a lokalni se opisnici raspodjeljuju u odgovarajuće grupe. Najčešće se koristi inicijalizacija slučajnim odabirom [102]. Algoritam je iterativan i odvija se u dva koraka [103]. U prvom se koraku obavlja procjena varijabli pridruživanja q_{tk} uz fiksne vrijednosti slikovnih riječi $\boldsymbol{\mu}_k$, dok se u drugom koraku obavlja procjena $\boldsymbol{\mu}_k$ uz fiksne vrijednosti varijabli pridruživanja:

$$\boldsymbol{\mu}_k = \frac{\sum_{t=1}^{T} q_{tk} \mathbf{x_t}}{\sum_{t=1}^{T} q_{tk}} \,. \tag{2.6}$$

U praksi se s obzirom na potencijalno veliki uzorak lokalnih opisnika $\mathbf{X} = \{\mathbf{x_t},\ t = 1 \dots T\}$ i razmjerno velik slikovni rječnik, koriste alternativni algoritmi kao što su Elkanov algoritam [104] ili algoritam približnih najbližih susjeda (engl. *Approximate Nearest Neighbor, ANN*) [105, 106].

Model raspodjele Gaussovih mješavina Model raspodjele Gaussovih mješavina (GMM) koristi se kao slikovni rječnik prilikom izgradnje Fisherovih vektora. Glavna prednost GMM-a jest da se uz proizvoljan broj Gaussovih komponenata, njime može aproksimirati bilo koja kontinuirana raspodjela [107].

Formalno, funkcija gustoće vjerojatnosti modela Gaussovih mješavina $\boldsymbol{\theta} = \{w_k, \boldsymbol{\mu}_k, \boldsymbol{\sigma}_k\}_{k=1}^{K}$,

$w_k \in \mathbb{R}$, $\boldsymbol{\mu}_k \in \mathbb{R}^D$, gdje $\boldsymbol{\sigma}_k^2 \in \mathbb{R}^D$ predstavlja dijagonalu matrice $\boldsymbol{\Sigma}_k \in \mathbb{R}^{D\times D}$ dana je izrazom:

$$p(\mathbf{x};\boldsymbol{\theta}) = \sum_{k=1}^{K} w_k \cdot p(\mathbf{x};\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k)\ . \tag{2.7}$$

Parametar K označava broj komponenata, dok w_k označava težinu miješanja, koja opisuje značaj komponente u okviru mješavine. Za težine miješanja vrijedi:

$$\forall k : w_k \geq 0, \quad \sum_{k=1}^{K} w_k = 1\ . \tag{2.8}$$

Kako bi se poštivalo gornje ograničenje, uvodi se reparametrizacija w_k primjenom normalizirane eksponencijalne funkcije (engl. *softmax*) koja implicitno zadovoljava gornje ograničenje [42, 108]:

$$w_k = \frac{\exp \alpha_k}{\sum_{j=1}^{K} \exp \alpha_j}, \quad \alpha_k \in \mathbb{R}\ . \tag{2.9}$$

Funkcija gustoće vjerojatnosti pojedine komponente definirana je izrazom:

$$p(\mathbf{x};\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k) = \frac{1}{(2\pi)^{D/2}|\boldsymbol{\Sigma}_k|^{1/2}} \exp\left\{ -\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_k)^T \boldsymbol{\Sigma}_k^{-1}(\mathbf{x}-\boldsymbol{\mu}_k) \right\}\ . \tag{2.10}$$

Parametar $\boldsymbol{\mu}_k \in \mathbb{R}^D$ označava srednju vrijednost raspodjele, a $\boldsymbol{\Sigma}_k \in \mathbb{R}^{D\times D}$ matricu kovarijacije. Parametar D označava dimenzionalnost lokalnih opisnika. Jednostavnosti radi, pretpostavlja se da su matrice kovarijacije dijagonalne, odnosno $\boldsymbol{\sigma}_k^2 \in \mathbb{R}^D$ označava dijagonalu matrice kovarijacije $\boldsymbol{\Sigma}_k$.

Parametri modela Gaussovih mješavina procjenjuju se algoritmom maksimizacije očekivanja (engl. *Expectation Maximization algorithm, EM*) [42] koji pripada paradigmi učenja bez nadzora.

2.1.5 Kôdiranje i sažimanje opisnika slikovnih okana

Kôdiranje opisnika slikovnih okana obavlja se u odnosu na slikovni rječnik (engl. *visual vocabulary, dictionary, codebook*) koji predstavlja statistiku općenitog sadržaja slike. Prilikom procesa kôdiranja, pojedini opisnik niske razine može biti pridružen isključivo jednoj slikovnoj riječi (najčešće najbližoj prema određenoj metrici). Opisani proces naziva se čvrsto pridruživanje (engl. *hard assignment*). S druge strane, kod mekog pridruživanja (engl. *soft assignment*) pojedini se lokalni opisnik može dodijeliti većem broju slikovnih riječi što u nekim slučajevima realnije opisuje stvarno činjenično stanje. U literaturi se mogu pronaći sljedeći pristupi kôdiranju lokalnih opisnika u odnosu na slikovni rječnik:

- kvantizacija vektora (engl. *vector quantization*): kôdiranje indeksom najbliže slikovne riječi (ili skupom indeksa k najbližih slikovnih riječi), primjerice kôdiranje u histogram

slikovnih riječi [40]

- kôdiranje na temelju kriterija *lokalnosti*, primjerice kôdiranje u vektor lokalno sažetih opisnika (engl. *Vector of Locally-Aggregated Descriptors, VLAD*) [45, 74]
- kôdiranje s obzirom na generativni model, primjerice kôdiranje na temelju Fisherove jezgre [43] rezultira Fisherovim vektorom [31, 33].

Nakon kôdiranja lokalnih opisnika nekim od navedenih postupaka, obavlja se operacija sažimanja (engl. *aggregation*) kôdiranih lokalnih opisnika u opisnik slike. Jedna od često korištenih operacija sažimanja odnosi se na sažimanje usrednjavanjem (engl. *sum pooling*) [31, 40], doprinosa lokalnih opisnika $\phi(\mathbf{x})$:

$$\phi(\mathbf{X}) = \frac{1}{T} \sum_{\mathbf{x} \in \mathbf{X}} \phi(\mathbf{x}) \,. \tag{2.11}$$

Parametar T označava kardinalnost skupa lokalnih opisnika. Uz sažimanje usrednjavanjem, u literaturi se mogu pronaći i radovi [109, 110] gdje se sažimanje obavlja na osnovu maksimalne vrijednosti (engl. *max pooling*) kôdiranih lokalnih opisnika:

$$\phi(\mathbf{X}) = \max_{\mathbf{x} \in \mathbf{X}} \phi(\mathbf{x}) \,. \tag{2.12}$$

Histogram slikovnih riječi

Prilikom kôdiranja lokalnih opisnika u histogram slikovnih riječi (engl. *Bag of Visual Words, BoVW*), primjenjuje se vektorska kvantizacija [40, 41]. Uz dani slikovni rječnik, gdje su slikovne riječi indeksirane vrijednostima $1 \ldots K$, vektorskom se kvantizacijom lokalnom opisniku $\mathbf{x} \in \mathbf{X}$ pridružuje najbliža slikovna riječ. Formalno, lokalni se opisnik kôdira jednojediničnim vektorom $\phi(\mathbf{x})$ (engl. *one-hot encoding*) čija je vrijednost jednaka ništici, izuzev dimenzije k koja označava indeks najbliže slikovne riječi:

$$\phi(\mathbf{x}) = [0 \ldots 1 \ldots 0] = \{\mathbf{1}_k(\mathbf{x})\}_{k=1:K} \,. \tag{2.13}$$

Izraz $\mathbf{1}_k(\mathbf{x}) : \mathbf{X} \to \{0,1\}$ označava indikatorsku funkciju, gdje

$$\mathbf{1}_k(\mathbf{x}) := \begin{cases} 1 & \text{akko } \mathbf{x} \mapsto k \\ 0 & \text{akko } \mathbf{x} \not\mapsto k \end{cases} . \tag{2.14}$$

Nakon operacije kôdiranja, na dobivene opisnike slikovnih okana $\phi(\mathbf{x})$ primjenjuje se operacija sažimanja usrednjavanjem što rezultira histogramom slikovnih riječi.

Unatoč jednostavnosti reprezentacije, vektorska kvantizacija i histogram slikovnih riječi imaju svoje nedostatke. Jedan od glavnih nedostataka odnosi se se na gubitak informacija prilikom vektorske kvantizacije [76, 111], gdje: 1) pojedini lokalni opisnik može dijeliti sličnosti s

Slika 2.6: Nedostaci histograma slikovnih riječi: primjeri raspodjele lokalnih opisnika dodijeljenih određenoj slikovnoj riječi. Broj lokalnih opisnika u oba slučaja (i lijevo i desno) je jednak, što rezultira jednakim histogramom slikovnih riječi. Međutim, raspodjela lokalnih opisnika u odnosu na centar grupe (velika masna točka) bitno je različita.

više slikovnih riječi ili 2) može jednako tako biti različit u odnosu na sve slikovne riječi. Posljednji je slučaj karakterističan za situacije gdje traženi objekti zauzimaju maleni udio površine slike. Prometni znakovi, čija se lokalizacija razmatra u odjeljku 5.2, primjer su objekta koji zauzima u prosjeku manje od 1 posto površine slike. Budući da se slikovni rječnik uči postupcima nenadziranog učenja na slučajno odabranom skupu opisnika iz pozitivnih i negativnih slika, za očekivati je da niti jedna slikovna riječ neće biti slična objektima tog razreda. U tom slučaju, vektorska kvantizacija i histogram slikovnih riječi nisu pogodno rješenje za reprezentaciju slike. Dodatno, dvije slike mogu rezultirati vrlo sličnim histogramima slikovnih riječi, gdje je broj lokalnih opisnika dodijeljenih pojedinoj slikovnoj riječi približan u oba slučaja, no raspodjela tih lokalnih opisnika u odnosu na centre grupa može biti bitno različita. Slika 2.6 ilustrira opisani problem za 2D slučaj.

Drugi nedostatak odnosi se na brzinu kôdiranja. Budući da histogrami slikovnih riječi bilježe isključivo broj opisnika dodijeljen određenoj slikovnoj riječi, kako bi se valjano opisala slika potreban je razmjerno velik slikovni rječnik. Proračun udaljenosti lokalnih opisnika u odnosu na veliki slikovni rječnik čini proces vektorske kvantizacije razmjerno skupim. Neka vrijede sljedeće oznake:

- N označava broj lokalnih opisnika u slici, gdje je $N \sim 10^4$ za gusto uzorkovane SIFT opisnike.
- K označava broj slikovnih riječi dobivenih algoritmom k-srednjih vrijednosti, gdje $K \sim 10^3$.
- D označava dimenzionalnost lokalnih opisnika, gdje $D \sim 10^2$ za SIFT opisnike.

Složenost izračuna histograma slikovnih riječi proporcionalna je $\mathrm{O}(NKD)$. Za gornji konkretan primjer vrijedi $\mathrm{O}(NKD) \sim 10^9$, odnosno potrebno je 10^9 operacija množenja kako bi se izračunao histogram slikovnih riječi.

Treći nedostatak histograma slikovnih riječi također je vezan uz veličinu slikovnog rječnika.

S obzirom na velik broj potrebnih slikovnih riječi, mnogi će odjeljci histograma ostati prazni, što upućuje na redundantnost reprezentacije.

Fisherov vektor

Fisherovi vektori koriste se za reprezentaciju slikovnih okana i slike u okviru ove disertacije te će stoga biti detaljnije opisani u odnosu na ostale opisnike porodice zbirke slikovnih riječi. Fisherovi vektori prvi su puta razmatrani za potrebe klasifikacije slika u okviru [31], a zasnivaju se na teorijski dobro utemeljenom konceptu Fisherove jezgre [43]. U nastavku odjeljka najprije je, stoga, dan opis Fisherove jezgre, a potom i Fisherovih vektora kao mehanizma za predstavljanje slikovnih okana i slika.

Fisherova jezgra U okviru ove disertacije, do sada su razmatrane vektorske reprezentacije slika i slikovnih okana koje omogućuju primjenu diskriminativnih modela za potrebe klasifikacije. Primjenom jezgrenih funkcija, podaci koji nemaju vektorski oblik mogu se predstaviti na način pogodan za linearnu klasifikaciju [1, 42]. Primjera radi, pretpostavimo da želimo obaviti klasifikaciju dviju slika različitih veličina, odnosno svakoj slici želimo pridijeliti oznaku razreda pozadina ili objekata uz pretpostavku binarne klasifikacije. Svaka se slika može promatrati kao skup slikovnih okana $\mathbf{X} = \{\mathbf{x_t},\, t = 1 \ldots T\}$. U tom slučaju, zadatak klasifikacije slika postaje zadatak odvajanja skupova slikovnih okana. Kôdiranje Fisherovim vektorima motivirano je upravo potrebom za klasifikacijom skupova podataka različitih veličina. Skupovi podataka koje se želi usporediti opisuju se generativnim modelom $p(\mathbf{X};\boldsymbol{\theta})$ parametriziranim s $\boldsymbol{\theta} \in \mathbb{R}^M$. Procjena parametara raspodjele $p(\mathbf{X};\boldsymbol{\theta})$ obavlja postupcima nenadziranog učenja (engl. *unsupervised learning*). Kako bi se dva skupa podataka $\mathbf{X}$ i $\mathbf{X}'$ načinila pogodnima za usporedbu, na temelju generativnog modela izvodi se Fisherov odziv (engl. *Fisher score*) $g(\mathbf{X};\boldsymbol{\theta}) \in \mathbb{R}^M$ za svaki skup podataka [43]:

$$g(\mathbf{X};\boldsymbol{\theta}) = \nabla_{\boldsymbol{\theta}} \log p(\mathbf{X};\boldsymbol{\theta})\,. \tag{2.15}$$

Fisherov odziv prema izrazu (2.15) definiran je kao gradijent logaritma izglednosti generativnog modela s obzirom na parametre $\boldsymbol{\theta}$. Intuitivno, Fisherov odziv pokazuje smjer promjene parametara generativnog modela $\boldsymbol{\theta}$ kako bi se što bolje opisali podaci. Dimenzionalnost Fisherova odziva $g(\mathbf{X};\boldsymbol{\theta})$ ne zavisi o veličini skupa $\mathbf{X}$, već isključivo o dimenzionalnosti parametara modela M. Drugim riječima, izračunom Fisherovih odziva dva se skupa podataka potencijalno različitih veličina preslikavaju u vektorsku reprezentaciju fiksne veličine. Uz pretpostavku da su opisnici slikovnih okana $\mathbf{x}_t$, koji čine skup $\mathbf{X}$, nezavisni i jednoliko raspodijeljeni (engl. *independent and identically distributed, i.i.d.*), Fisherov odziv skupa $\mathbf{X}$ može se izračunati usrednjavanjem

Fisherovih odziva pojedinih okana *:

$$g(\mathbf{X};\boldsymbol{\theta}) = \frac{1}{T}\sum_t \nabla_{\boldsymbol{\theta}} \log p(\mathbf{x}_t;\boldsymbol{\theta}) = \frac{1}{T}\sum_t g(\mathbf{x}_t;\boldsymbol{\theta}) \,. \tag{2.16}$$

Uz definirane Fisherove odzive, skupovi podataka $\mathbf{X}$ i $\mathbf{X}'$ mogu se usporediti primjenom jezgrene funkcije koja predstavlja skalirani umnožak između dvaju Fisherovih odziva:

$$K(\mathbf{X},\mathbf{X}') = g(\mathbf{X};\boldsymbol{\theta})^T \mathbf{F}(\boldsymbol{\theta})^{-1}\, g(\mathbf{X}';\boldsymbol{\theta}) \,. \tag{2.17}$$

Izraz (2.17) označava Fisherovu jezgru, gdje $\mathbf{F}(\boldsymbol{\theta}) \in \mathbb{R}^{M\times M}$ označava Fisherovu informacijsku matricu (*FIM*):

$$\mathbf{F}(\boldsymbol{\theta}) = \mathbb{E}_{\mathbf{X}\sim p(\mathbf{X};\boldsymbol{\theta})}\left[g(\mathbf{X};\boldsymbol{\theta}) g(\mathbf{X};\boldsymbol{\theta})^T\right] \,. \tag{2.18}$$

Fisherova informacijska matrica $\mathbf{F}(\boldsymbol{\theta})$ predstavlja matricu kovarijacije Fisherovih odziva, a njezina je uloga načiniti Fisherovu jezgru neosjetljivom na reparametrizaciju modela $\boldsymbol{\theta} \to \psi(\boldsymbol{\theta})$ [1]. Proračun Fisherove informacijske matrice direktno prema (2.18) često nije moguć sa stanovišta potrošnje računalnih resursa. Iz tog razloga, u praksi se koriste različite aproksimacije, primjerice srednjom vrijednosti na određenom skupu primjera [42] ili analitičkom aproksimacijom kao što je predloženo u [31]. Važno svojstvo Fisherove informacijske matrice jest da je pozitivno definitna, odnosno vrijedi [1]:

$$\boldsymbol{\alpha}^T \cdot \mathbf{F}(\boldsymbol{\theta}) \cdot \boldsymbol{\alpha} = \mathbb{E}_{\mathbf{X}\sim p(\mathbf{X};\boldsymbol{\theta})}\left[(g(\mathbf{X};\boldsymbol{\theta})^T \boldsymbol{\alpha})^2\right] > 0 \,. \tag{2.19}$$

Parametar $\boldsymbol{\alpha}$ označava vektor različit od nul vektora, $\boldsymbol{\alpha} \neq \mathbf{0}$. Posljedica opisanog svojstva jest da se može obaviti Cholesky dekompozicija [112] njezina inverza $\mathbf{F}(\boldsymbol{\theta})^{-1}$:

$$\mathbf{F}(\boldsymbol{\theta})^{-1} = \mathbf{L}(\boldsymbol{\theta})^T \mathbf{L}(\boldsymbol{\theta}) \,. \tag{2.20}$$

Primjenom (2.20), Fisherova se jezgra može prikazati eksplicitno preko skalarnog produkta funkcija preslikavanja $\Phi_{\boldsymbol{\theta}}(\mathbf{x})$:

$$K(\mathbf{X},\mathbf{X}') = g(\mathbf{X};\boldsymbol{\theta})^T \mathbf{F}(\boldsymbol{\theta})^{-1}\, g(\mathbf{X}';\boldsymbol{\theta}) = \phi_{\boldsymbol{\theta}}(\mathbf{X})^T \phi_{\boldsymbol{\theta}}(\mathbf{X}') \,. \tag{2.21}$$

Dobivena funkcija preslikavanja $\Phi_{\boldsymbol{\theta}}(\mathbf{X})$ naziva se Fisherov vektor [31, 33]:

$$\Phi_{\boldsymbol{\theta}}(\mathbf{X}) = \mathbf{L}(\boldsymbol{\theta}) \cdot g(\mathbf{X};\boldsymbol{\theta}) \,. \tag{2.22}$$

*Sasvim općenito, izglednost skupa nezavisnih jednoliko raspodijeljenih podataka $\mathbf{X} = \{\mathbf{x_t},\, t = 1\ldots T\}$ odgovara $p(\mathbf{X};\boldsymbol{\theta}) = \prod_{t=1}^{T} p(\mathbf{x}_t;\boldsymbol{\theta})$. Jednostavnosti radi, često se koristi logaritam izglednosti: $\log p(\mathbf{X};\boldsymbol{\theta}) = \sum_{t=1}^{T} \log p(\mathbf{x}_t;\boldsymbol{\theta})$. Fisherov odziv predstavlja gradijent logaritma izglednosti pa se stoga Fisherov odziv skupa $\mathbf{X}$ može prikazati kao prosječna vrijednost Fisherovih odziva $\mathbf{x}_t$.

Dimenzionalnost dobivena Fisherova vektora $\Phi_{\boldsymbol{\theta}}(\mathbf{X})$ odgovara dimenzionalnosti Fisherova odziva $g(\mathbf{X};\boldsymbol{\theta})$, dakle ne ovisi o dimenzionalnosti $\mathbf{X}$. Fisherova jezgra jedna je od rijetkih jezgrenih funkcija gdje je funkciju preslikavanja $\Phi_{\boldsymbol{\theta}}(\mathbf{X})$ moguće izračunati izravno na temelju značajki u originalnom prostoru [42]. Klasifikacija skupova podataka $\mathbf{X}$ i $\mathbf{X}'$ tada je moguća pomoću linearnog diskriminativnog modela učenog na temelju pripadajućih Fisherovih vektora $\Phi_{\boldsymbol{\theta}}(\mathbf{X})$ i $\Phi_{\boldsymbol{\theta}}(\mathbf{X}')$. Drugim riječima, reprezentacija Fisherovim vektorima omogućava objedinjavanje prednosti generativnih i diskriminativnih modela. Klasifikacija linearnim modelom učenim nad Fisherovim vektorima ekvivalentna je klasifikaciji sa nelinearnom Fisherovom jezgrom. Prednost ugrađivanja u prostor Fisherovih vektora u odnosu na klasifikaciju Fisherovom jezgrom leži u činjenici da su linearni modeli znatno jednostavniji za učenje. Naime, u slučaju klasifikacije Fisherovom jezgrom, prilikom učenja se obavlja računski skup proračun inverza Gramove matrice dimenzija $N \times N$, gdje N označava broj uzoraka za učenje [42].

Fisherovi vektori u analizi slike Nakon formalne definicije Fisherovih vektora, u ovom se odjeljku razmatra primjena u analizi slike. U aspektu analize slike, skup vektora $\mathbf{X} = \{\mathbf{x_t},\, t = 1 \ldots T\}$, $\mathbf{x_t} \in \mathbb{R}^D$ označava skup lokalnih opisnika slikovnih okana dobivenih nekim od postupka uzorkovanja opisanim u odjeljku 2.1.1. Pretpostavlja se da su lokalni opisnici nezavisno i jednoliko raspodijeljeni (engl. *independent and identically distributed, i.i.d.*). U okviru eksperimentalnog vrednovanja u poglavlju 5 koriste se SIFT opisnici [37] i konvolucijske značajke [39], ali to mogu biti i drugi opisnici niske razine opisani u odjeljku 2.1.2.

Kao generativni model kojim se opisuje skup $\mathbf{X} = \{\mathbf{x_t},\, t = 1 \ldots T\}$, najčešće se koristi model mješavine Gaussovih raspodjela GMM parametriziran s $\boldsymbol{\theta} = \{\alpha_k, \boldsymbol{\mu}_k, \boldsymbol{\sigma}_k\}_{k=1}^K$, $\alpha_k \in \mathbb{R}$, $\boldsymbol{\mu}_k \in \mathbb{R}^D$, $\boldsymbol{\sigma}_k \in \mathbb{R}^D$ opisan u okviru odjeljka 2.1.4. Shodno tome, vjerojatnost da k-ta komponenta GMM modela generira opisnik $\mathbf{x}$ dana je izrazom:

$$p(k|\mathbf{x}) = \frac{p(k) \cdot p(\mathbf{x};k)}{p(\mathbf{x})} = \frac{w_k \cdot p(\mathbf{x};\boldsymbol{\mu}_k,\boldsymbol{\sigma}_k)}{p(\mathbf{x};\boldsymbol{\theta})} = \frac{w_k \cdot p(\mathbf{x};\boldsymbol{\mu}_k,\boldsymbol{\sigma}_k)}{\sum_{i=1}^K w_i \cdot p(\mathbf{x};\boldsymbol{\mu}_i,\boldsymbol{\sigma}_i)}\,. \tag{2.23}$$

Izraz w_k označava težinu odgovarajuće komponente Gaussove mješavine, a definiran je jednadžbom (2.9). U odnosu na histogram slikovnih riječi, gdje se pojedini lokalni opisnik dodjeljuje isključivo najbližoj slikovnoj riječi, izrazom (2.23) definirano je meko pridruživanje (engl. *soft-assignment*) opisnika $\mathbf{x}$ u odnosu na komponente slikovnog rječnika realiziranog GMM-om.

Primjenom modela mješavine Gaussovih raspodjela u izrazima (2.22) i (2.15), gradijenti logaritma funkcije izglednosti $p(\mathbf{x};\boldsymbol{\theta})$ u odnosu na parametre $\alpha_k, \boldsymbol{\mu}_k, \boldsymbol{\sigma}_k$ definirani su preko [33]:

$$\phi_{\alpha_k}(\mathbf{x}) = \frac{p(k|\mathbf{x}) - w_k}{\sqrt{w_k}}\,. \tag{2.24}$$

$$\phi_{\boldsymbol{\mu}_k}(\mathbf{x}) = \frac{p(k|\mathbf{x})}{\sqrt{w_k}} \cdot \frac{\mathbf{x} - \boldsymbol{\mu}_k}{\boldsymbol{\sigma}_k}\,. \tag{2.25}$$

$$\phi_{\boldsymbol{\sigma}_k}(\mathbf{x}) = \frac{p(k|\mathbf{x})}{\sqrt{2w_k}} \cdot \left[\frac{(\mathbf{x}-\boldsymbol{\mu}_k)^2}{\boldsymbol{\sigma}_k^2} - 1 \right] . \tag{2.26}$$

Konačna vektorska reprezentacija dobiva se konkateniranjem izraza $\phi_{\alpha_k}(\mathbf{x})$, $\phi_{\boldsymbol{\mu}_k}(\mathbf{x})$ i $\phi_{\boldsymbol{\sigma}_k}(\mathbf{x})$ za sve komponente $k = 1 \ldots K$ u jedinstveni vektor $\phi_{\boldsymbol{\theta}}(\mathbf{x})$, čija je dimenzionalnost jednaka $K(2D+1)$.

Budući da se Fisherovi vektori dobivaju normalizacijom Fisherovih odziva Fisherovom informacijskom matricom, Fisherov vektor slike također se može izračunati usrednjavanjem Fisherovih vektora slikovnih okana [33]:

$$\Phi_{\boldsymbol{\theta}}(\mathbf{X}) = \frac{1}{T} \sum_{t=1}^{T} \phi_{\boldsymbol{\theta}}(\mathbf{x_t}) . \tag{2.27}$$

Opisano svojstvo Fisherovih vektora naziva se *aditivnost* i predstavlja ključan element sustava za lokalizaciju odozdo prema gore. Nadalje, očekivanje Fisherova odziva (pa tako i Fisherova vektora) jednako je nul vektoru:

$$\mathbb{E}_{\mathbf{x} \sim p(\mathbf{x};\boldsymbol{\theta})}[\phi(\mathbf{x};\boldsymbol{\theta})] = \mathbf{0} . \tag{2.28}$$

Shodno navedenom svojstvu isčezavanja očekivanja, reprezentacija slike Fisherovim vektorima poništava se utjecaj pozadinske informacije neovisne o slici [33]. Drugim riječima slika je opisana upravo onim značajkama koje ju čine drugačijima u odnosu na globalnu statistiku sadržaja slike predstavljenu slikovnim rječnikom. Matrica kovarijacije je jednaka jediničnoj matrici:

$$\mathbb{E}_{\mathbf{x} \sim p(\mathbf{x};\boldsymbol{\theta})}[\phi(\mathbf{x};\boldsymbol{\theta})\ \phi(\mathbf{x};\boldsymbol{\theta})^T] = \mathbf{I} . \tag{2.29}$$

Opisano svojstvo posljedica je normalizacije s $\mathbf{L}(\boldsymbol{\theta}) = \mathbf{F}(\boldsymbol{\theta})^{\frac{-1}{2}}$, gdje Fisherova informacijska matrica $\mathbf{F}(\boldsymbol{\theta})$ uslijed gornjeg svojstva iščezavanja očekivanja (2.28) predstavlja matricu kovarijacije Fisherova odziva.

Vektor lokalno sažetih opisnika VLAD

Kôdiranje u vektor lokalno sažetih opisnika (engl. *Vector of Locally-Aggregated Descriptors, VLAD*) predloženo je u okviru rada [74]. Slično kao i u slučaju histograma slikovnih riječi, kao slikovni rječnik koriste se centri grupa dobiveni grupiranjem k-srednjih vrijednosti, a pojedini lokalni opisnik dodjeljuje se isključivo najbližoj slikovnoj riječi. Proces kôdiranja opisnika slike $\mathbf{X}$ obavlja se prema tzv. kriteriju *lokalnosti*, gdje se za svaku slikovnu riječ $\boldsymbol{\mu}_k$ akumuliraju

razlike lokalnih opisnika $\mathbf{x_i}$ dodijeljenih toj slikovnoj riječi u odnosu na centar grupe:

$$\phi(\mathbf{X})_k = \sum_{\mathbf{x_i} \in \mathbf{X} \sim \boldsymbol{\mu}_k} \mathbf{x_i} - \boldsymbol{\mu}_k \,. \tag{2.30}$$

Primjenom izraza (2.30) opisuje se raspodjela lokalnih opisnika u odnosu na centar grupe. Konačan opisnik dobiva se konkatenacijom doprinosa pojedinih slikovnih riječi:

$$\phi(\mathbf{X}) = [\phi(\mathbf{X})_1 \ldots \phi(\mathbf{X})_k \ldots \phi(\mathbf{X})_K] \,. \tag{2.31}$$

Uz lokalne opisnike dimenzije D i K slikovnih riječi, konačna dimenzionalnost vektora odgovara $K \cdot D$.

Rasprava postupaka kôdiranja

U okviru ovog odjeljka razmatraju se prednosti i razlike Fisherovih vektora u odnosu na ostale opisnike porodice zbirke slikovnih riječi. Tablica 2.1 zorno prikazuje usporedbu različitih opisnika s obzirom na izvedbu slikovnog rječnika, način pridruživanja lokalnih opisnika u odnosu na slikovni rječnik, dimezionalnost opisnika i konačno način kôdiranja, odnosno doprinos k-te slikovne riječi konačnom opisniku slikovnog okna. Tablica pokazuje da Fisherovi vektori (FV) predstavljaju proširenje histograma slikovnih riječi (BoVW), dok vektori lokalno sažetih opisnika (VLAD) predstavljaju pojednostavljenje Fisherovih vektora.

U odnosu na reprezentaciju histogramima slikovnih riječi, Fisherovi vektori unose nekoliko proširenja. Prva razlika odnosi se na izbor slikovnog rječnika, gdje Fisherovi vektori koriste generativni model kao slikovni rječnik. Slikovni rječnik predstavlja generalnu statistiku sadržaja slike. Što slikovni rječnik bolje opisuje podatke, samo kôdiranje u odnosu na slikovni rječnik je vjerodostojnije. Kod Fisherovih vektora, svaka slikovna riječ opisana je Gaussovom raspodjelom, odnosno pripadajućom srednjom vrijednosti i varijancom. Kod slikovnog rječnika u slučaju histograma slikovnih riječi ta se informacija gubi. Nadalje, Fisherovi vektori bilježe više informacija u odnosu na histograme slikovnih riječi. Fisherovi vektori bilježe gradijente logaritamske izglednosti u odnosu na težinu miješanja, srednju vrijednost i varijancu pojedine slikovne riječi (GMM komponente), dok histogram slikovnih riječi bilježi isključivo broj lokalnih opisnika dodijeljenih slikovnoj riječi. Shodno tome, Fisherovi vektori zahtijevaju znatno manji slikovni rječnik, što znatno utječe na ubrzanje procesa kôdiranja. Naposljetku, kôdiranje Fisherovim vektorima uključuje nelinearne transformacije. Kako bi se odvojili uzorci u prostoru histograma slikovnih riječi [40], obično se koristi nelinearni klasifikator poput stroja s potpornim vektorima s nelinearno jezgrenom funkcijom (engl. *kernel trick*) [42] (radijalne bazne funkcije, polinomijalna jezgra, sigmoidna jezgra i druge). Budući da je transformacija u prostor Fisherovih vektora temeljena na Fisherovoj jezgri, uzorci u prostoru Fisherovih vek-

Tablica 2.1: Usporedba opisnika porodice zbirke slikovnih riječi: histograma slikovnih riječi (BoVW), vektora lokalno sažetih opisnika (VLAD) i Fisherovih vektora (FV).

Opisnik	Sl. rječnik	Pridruž.	Dimenzija	Kôdiranje $\phi(\mathbf{x})_k$
BoVW [40]	*k-means* centri	čvrsto	$K \cdot D$	$\mathbf{1}_k(\mathbf{x})$
VLAD [74]	*k-means* centri	čvrsto	$K \cdot D$	$\mathbf{x} - \boldsymbol{\mu}_k$
FV [33]	GMM	meko	$2 \cdot (K \cdot D + 1)$	$\left[\frac{p(k\|\mathbf{x})-w_k}{\sqrt{w_k}} \; \frac{p(k\|\mathbf{x})}{\sqrt{w_k}} \frac{\mathbf{x}-\boldsymbol{\mu}_k}{\boldsymbol{\sigma}_k} \; \frac{p(k\|\mathbf{x})}{\sqrt{2w_k}} \left(\frac{(\mathbf{x}-\boldsymbol{\mu}_k)^2}{\boldsymbol{\sigma}_k^2} - 1\right)\right]$

tora pogodni su za odvajanje linearnim klasifikatorima. Primjena linearnih klasifikatora putem skalarnog produkta znatno je jednostavnija i efikasnija u odnosu na izračun jezgrene funkcije. Drugim riječima, Fisherovi vektori omogućavaju brži proračun odziva slikovnih regija.

Usporedba Fisherovih vektora i vektora lokalno sažetih opisnika razmatrana je u okviru rada [113]. Jegou i dr. u [113] pokazuju da se vektor lokalno sažetih opisnika može promatrati kao neprobabilistička inačica Fisherovih vektora uz sljedeće aproksimacije: 1) meko pridruživanje u slučaju Fisherovih vektora zamjenjuje se čvrstim, 2) razmatra se odnos lokalnog opisnika i slikovne riječi isključivo na temelju srednje vrijednosti.

2.2 Reprezentacije prostornog rasporeda slike

Primjenom vektorskih reprezentacija slike iz porodice zbirke slikovnih riječi ne bilježe se razlike u prostornim koordinatama i mjerilima na kojima su izdvojeni lokalni opisnici. Prostorni se raspored bilježi isključivo unutar slikovnih okana odgovarajućim lokalnim opisnicima. S druge strane, pokazuje se da modeliranje prostornih odnosa između slikovnih okana može doprinijeti poboljšanju učinkovitosti zadataka klasifikacije slika i lokalizacije objekata [6, 75, 114, 115]. U nastavku odjeljka dan je pregled recentnih radova na području predstavljanja prostornog rasporeda slike.

Sasvim općenito, modeli reprezentacije prostornog rasporeda mogu se podijeliti u dvije osnovne skupine s obzirom na opseg prostornog rasporeda koji se njima predstavlja:

- prostorni opisnici za modeliranje globalnih prostornih odnosa u okviru slike [32, 75, 97, 108, 116, 117, 118, 119]
- prostorni opisnici za predstavljanje lokalnih prostornih struktura [114, 115, 120, 121, 122, 123].

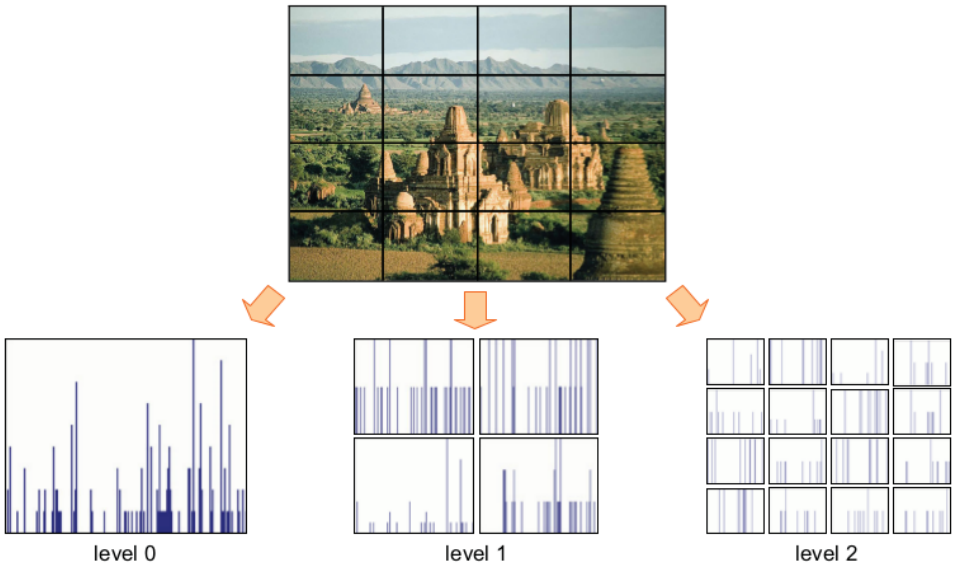
2.2.1 Opisnici za predstavljanje globalnog prostornog rasporeda

Koncept globalnog prostornog rasporeda odnosi se na pojavljivanje pojedinih slikovnih elemenata na točno određenim pozicijama u slici. Primjerice, autori u [47] navode kako scene

eksterijera najčešće poprimaju karakterističan globalni prostorni raspored, gdje prometne scene uključuju različite građevine s lijeve i desne strane, a u središtu cestu. Dodatno, objekti specifičnog razreda također mogu poprimati karakteristične pozicije u okviru slike. Jedan od primjera globalnog rasporeda vezanog uz određeni razred jest razred automobila, gdje se automobili u slici najčešće pojavljuju na cesti. U nastavku je dan opis značajnih radova kojima se opisuju reprezentacije globalnog prostornog rasporeda.

Najčešće primjenjivana metoda reprezentacije globalnog prostornog rasporeda jest piramidalna prostorna reprezentacija (engl. *Spatial Pyramid Matching, SPM*) [75]. Slika 2.7 ilustrira shematski prikaz metode prostorne piramide temeljene na histogramu slikovnih riječi (engl. *Bag of Visual Words*) [40, 41]. Postupak se zasniva na iterativnoj podijeli slike u više razina. U svakoj se narednoj razini, slika dijeli u sve finije i finije ćelije te je u i-toj razini piramide slika podijeljena u $2^i \times 2^i$ ćelija. Za svaku se ćeliju gradi odgovarajući histogram slikovnih riječi. Prostorne se piramide mogu uparivati (engl. *matching*) primjenom piramidalnih jezgri gdje se značajkama na višim razinama piramide dodjeljuju veći iznosi težina. Glavna prednost reprezentacije prostornom piramidom jest da se relativno jednostavno može integrirati s gotovo svakim opisnikom iz porodice opisnika zbirke slikovnih riječi [6, 32, 116]. Glavni nedostaci modela prostorne piramide su sljedeći: 1) organizacija i podjela u prostorne ćelije nije nužno optimalna, 2) povećava se dimenzionalnost opisnika slike čime se povećava vjerojatnost pojave prenaučenosti (engl. *overfitting*) [34, 124]. U literaturi se, stoga, mogu pronaći brojni radovi u kojima se razmatraju poboljšanja reprezentacije prostornom piramidom. Autori u [97] razmatraju alternativne mogućnosti podijele slike u ćelije na svakoj razini. Tako u nultoj razini raspolažu čitavom slikom kao jedinstvenom ćelijom (kao i originalna SPM reprezentacija), u prvoj razini dijele sliku u tri ćelije 1×3 (gornju, središnju i donju ćeliju), dok je u trećoj razini slika podijeljena u 2×2 prostorne ćelije. Nadalje, autori u [119] predlažu proširenje SPM modela učenjem specifičnih težina za svaku razinu piramide na skupu podataka za validaciju. U radu predstavljenom u [118], predlaže se učenje specifičnih SPM modela za svaki razred, pri čemu se iterativno dodaju prostorne ćelije tijekom učenja.

Radovi [108, 117] razmatraju modeliranje globalnog prostornog rasporeda kao proširenja reprezentacije Fisherovih vektora. Oba pristupa zaobilaze potencijalno problematično particioniranje u prostorne ćelije nužno za SPM reprezentaciju. Autori u [117] proširuju Fisherov vektor slike pozicijom u okviru koordinatnog sustava slike (x, y, m), gdje x i y označavaju poziciju po apscisi i ordinati, dok m označava mjerilo na kojem je značajka izdvojena. S druge strane, autori u [108] predstavljaju prostorni raspored slikovnih okana dodijeljenih određenoj slikovnoj riječi pomoću modela mješavine Gaussovih raspodjela. Pri tome se uvodi aproksimacija čvrstim pridruživanjem, gdje je svako slikovno okno dominantno dodijeljeno isključivo jednoj slikovnoj riječi. Umjesto direktnog proširivanja Fisherova vektora prostornim koordinatama kao u [117], prostorna se reprezentacija dobiva kôdiranjem koordinata u odnosu na odgovarajući prostorni



Slika 2.7: Primjer shematskog prikaza reprezentacije prostornom piramidom (engl. *Spatial Pyramid*) histograme slikovnih riječi (engl. *Bag of Visual Words, BoVW*) [75]. Na prvoj razini piramide (level 1), slika se dijeli u četiri ćelije, a potom se za svaki od ćelija gradi histogram slikovnih riječi. Konačan prostorni opisnik dobiva se konkatenacijom opisnika pojedinih ćelija.

model mješavine Gaussovih raspodjela. Oba modela reprezentacije [108, 117] postižu sumjerljive rezultate u odnosu na model prostorne piramide, no uz značajno kompaktniju prostornu reprezentaciju.

2.2.2 Opisnici za predstavljanje lokalnog prostornog rasporeda

Modeliranje lokalnog prostornog rasporeda u pravilu je složenije zbog činjenica da se specifične lokalne strukture mogu 1) pojaviti bilo gdje u slici, a 2) mogu i izostati u pojedinim scenama. Primjerice, autori u [47] navode kako objekti karakteristični za scene interijera najčešće imaju specifičan lokalni prostorni raspored (kućanski se aparati mogu pojavljivati na proizvoljnim pozicijama u kuhinji).

U literaturi se mogu pronaći različiti pristupi predstavljanju lokalnog prostornog rasporeda podijeljeni u nekoliko skupina:

- pristupi temeljeni na modeliranju neuređenih pojavljivanja slikovnih riječi (engl. *co-occurrence of visual words*) [115, 120, 121, 125, 126]
- pristupi temeljeni na modeliranju prostornih konstelacija između slikovnih riječi [114, 122, 123]
- pristupi temeljeni na modeliranju prostornih odnosa kao opisnika druge razine, pri čemu se najprije određuju približne lokacije objekata, a reprezentacije prostornog rasporeda koriste se za poboljšanje klasifikacijske i lokalizacijske učinkovitosti [123, 127, 128].

Pristupi temeljeni na modeliranju neuređenih pojavljivanja slikovnih riječi Pristup predstavljen u [120] zasniva se na izgradnji hijerarhijskih značajki srednje razine (engl. *mid-level features*). Postupak je iterativan pri čemu se na svakoj razini hijerarhije računaju histogrami slikovnih riječi za progresivno sve veća lokalna okruženja. Opisnici dobiveni na višim razinama hijerarhije implicitno kôdiraju odnose među slikovnim riječima. Autori u [115] modeliranju globalne i lokalne prostorne odnose izdvajanjem značajki srednje razine (tzv. *correlograms*) kojima mjere korelaciju pojavljivanja parova slikovnih riječi u okviru unaprijed definiranih lokalnih susjedstva. Nadalje, autori u [121] predlažu reprezentaciju značajkama srednje razine za svaku slikovnu riječ na temelju odgovarajućeg lokalnog konteksta. Predložena reprezentacija poprima oblik histograma u odnosu na slikovne riječi dodijeljene oknima u okviru lokalnog konteksta. Problem modeliranja međusobnih pojavljivanja slikovnih riječi također se razmatra tehnikama dubinske analize podataka (engl. *data mining*) [125, 129], konkretno tehnikom analize čestih skupova (engl. *frequent itemset mining*). U tom se aspektu BoVW histogrami promatraju kao transakcije, a međusobna zajednička pojavljivanja slikovnih riječi kao česti uzorci [126] ili skupovi [125, 129].

Pristupi temeljeni na modeliranju prostornih konstelacija između slikovnih riječi Primjeri radova temeljeni na modeliranju prostornih konstelacija između slikovnih riječi uključuju [114, 122, 123]. U okviru [114], autori koriste histogram za modeliranje prostornih odnosa među parovima slikovnih riječi. Glavni problem ovog pristupa je stabilnost. Budući da se nekolicina slikovnih riječi manifestira u bilo kojoj diskriminativnoj regiji slike (tzv. reprezentativne slikovne riječi), potreban je razmjerno velik broj parova slikovnih riječi da bi se opisalo pojedino obilježje određenog razreda. Dodatno, zbog primjene histograma, opisani pristup zahtjeva potencijalno velik broj primjera za učenje kako bi se korektno načinila diskretizacija histograma u odjeljke. S druge strane, pristup predstavljen u [122] odabire relativno malen slikovni rječnik od stotinjak slikovnih riječi te razmatra jednostavne prostorne odnose prema blizini i orijentaciji. Rad [123] predstavlja iterativan postupak probira diskriminativnih slikovnih riječi. U svakoj se iteraciji za svaku slikovnu riječ uči diskriminativni model (stroj sa potpornim vektorima, (engl. *Support Vector Machine, SVM*)) na podskupu podataka. Model se primjenjuje na drugom podskupu podataka, a pozitivni odzivi se grupiraju kako bi se formirale slikovne riječi za narednu iteraciju. U fazi postprocesiranja obavlja se pretraživanje parova prostorno koreliranih slikovnih riječi čime se povećava učinkovitost klasifikacije.

Pristupi temeljeni na modeliranju prostornih odnosa kao opisnika druge razine Primjeri radova koji se temelje na modeliranju prostornih odnosa kao opisnika druge razine uključuju [127, 128] i već opisani pristup [123]. Ideja navedenih pristupa jest pronaći približne lokacije objekata [127] ili regija koje potencijalno sadrže objekte [123, 128]. Dobivene regije koriste

se kao ćelije za prostorno sažimanje (engl. *spatial aggregation*) [127], a mogu se koristiti kao značajke srednje razine za konstrukciju reprezentacije više razine [123].

2.2.3 Rasprava reprezentacija prostornog rasporeda

U okviru ove disertacije, razmatra se problem predstavljanja lokalnog prostornog rasporeda u svrhu povećanja lokalizacijske učinkovitosti. Modeli reprezentacije prostornog rasporeda primjenjuju se kao opisnici druge razine za predstavljanje lokalnih okruženja slikovnih okana za koja lokalizacijski model prve razine daje pozitivan odziv. U tome aspektu, predloženi pristup sličan je radovima [123, 127, 128] budući da se primjenjuju na slikovne regije koje potencijalno sadrže objekte od interesa. Slično kao i u radovima predstavljenim u [114, 122, 123], prostornim opisnicima predloženim u okviru ove disertacije predstavljaju se prostorne konstelacije odnosa između parova slikovnih riječi. U odnosu na model predstavljen u [114], koji pati od problema stabilnosti uslijed velikog broja mogućih parova slikovnih riječi, u okviru ove disertacije stabilnost se postiže rijetkim lokalizacijskim modelima prve razine. Rijetki modeli identificiraju diskriminativne slikovne riječi, a prostorni se raspored gradi isključivo za parove diskriminativnih slikovnih riječi. Eksperimenti u poglavlju 5 pokazuju da je u slučajevima gdje se postiže najbolja lokalizacijska učinkovitost, tek oko 1 posto slikovnih riječi diskriminativno (tablice 5.4 i 5.2). S druge strane, predložena reprezentacija prostornim Fisherovim vektorima slična je i pristupu objavljenom u okviru rada [108] gdje su prostorni opisnici također zasnovani na konceptu Fisherove jezgre. Razlika između predloženog pristupa i pristupa opisanog u [108] je u tome da se u [108] opisuje globalni prostorni raspored te se umjesto prostornih odnosa parova slikovnih riječi modelira raspored slikovnih okana dodijeljenih određenoj slikovnoj riječi.

2.3 Lokalizacija objekata u slikama

Zadatak lokalizacije objekata jest odrediti točnu lokaciju objekta u smislu pozicije u okviru koordinatnog sustava slike te pripadajućeg opisanog poligona. Velik broj radova [4, 130, 131] pristupa tom problemu na način da iz slike izdvaja velik broj slikovnih regija na koje se primjenjuju binarni lokalizacijski modeli učeni zasebno za pojedini razred objekata. Slikovne regije maksimalnog odziva određuju potencijalne lokacije objekata. S obzirom na potencijalno velik broj mogućih hipoteza, gdje broj potencijalnih lokacija raste kvadratno u odnosu na broj piksela slike [132], računski je neučinkovito računati odziv za svaku regiju dobivenu iscrpnim pretraživanjem. Kako bi se vrijeme izvođenja približilo izvođenju u stvarnom vremenu, potrebno je pronaći pametnu lokalizacijsku strategiju. Pojam lokalizacijske strategije uključuje algoritme pretraživanja potencijalnih lokacija objekata, ali i metode reprezentacije slike koje omogućuju

efikasan proračun odziva lokalizacijskog modela za pojedinu slikovnu regiju (primjerice, integralna slika [3, 115, 133]). U nastavku su navedeni najznačajniji primjeri lokalizacijskih strategija dostupnih u literaturi:

- lokalizacija primjenom ulančanog klasifikatora, gdje svaki stupanj lanca odgovara ojačanom klasifikatoru (engl. *cascade of boosted classifiers*) [3]
- učinkovito pretraživanje potprozora (engl. *Efficient Subwindow Search, ESS*) [132]
- postupak pretraživanja zasnovan na naučenim lokalizacijskim prijedlozima na temelju mjere *objektnosti* (engl. *objectness*) [134]
- postupak pretraživanja temeljen na segmentaciji superpikselima [44].

U okviru narednih odjeljaka, detaljno su opisane navedene lokalizacijske strategije, pripadajuća poboljšanja i primjene u računalnom vidu. Uz detaljan opis lokalizacijskih strategija, razmatra se i usporedba u odnosu na lokalizacijski pristup Fisherovim vektorima i rijetkim modelima predstavljen u okviru ove disertacije. Usporedba se obavlja u terminima računske učinkovitosti i potpunosti algoritma pretraživanja, gdje neke od strategija ne obavljaju globalno optimalno pretraživanje prostora hipoteza.

2.3.1 Lokalizacija kaskadom ojačanih klasifikatora

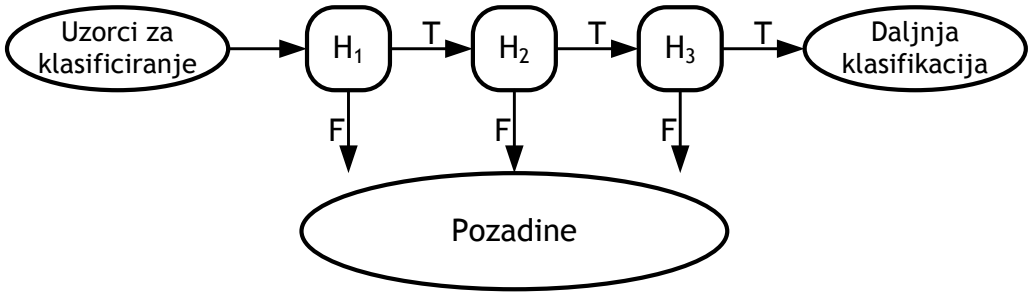
Strategija lokalizacije ulančavanjem progresivno sve složenijih lokalizacijskih modela za potrebe detekcije ljudskih lica predstavljena je u okviru [3]. Primjer ulančanog klasifikatora ilustriran je u okviru slike 2.8. Prva razina kaskade odgovara modelu sa stopostotnim odzivom i preciznošću od tek pedesetak posto, gdje je ideja da se inicijalno eliminira što više pozadinskih prozora. Svaka naredna razina primjenjuje se isključivo nad prozorima koje je prethodna razina označila kao pozitivne čime se postiže učinkovito pretraživanje značajki.

Opisani pristup lokalizaciji temelji se na modelu reprezentacije integralnom slikom [133, 135, 136, 137]. Reprezentacija integralnom slikom omogućuje učinkovit proračun vrijednosti lokalnih opisnika ilustriran u okviru slike 2.9. Vrijednost integralne slike ii na poziciji $ii(x,y)$ odgovara sumi vrijednosti intenziteta piksela iznad i lijevo u odnosu na danu poziciju u originalnoj slici:

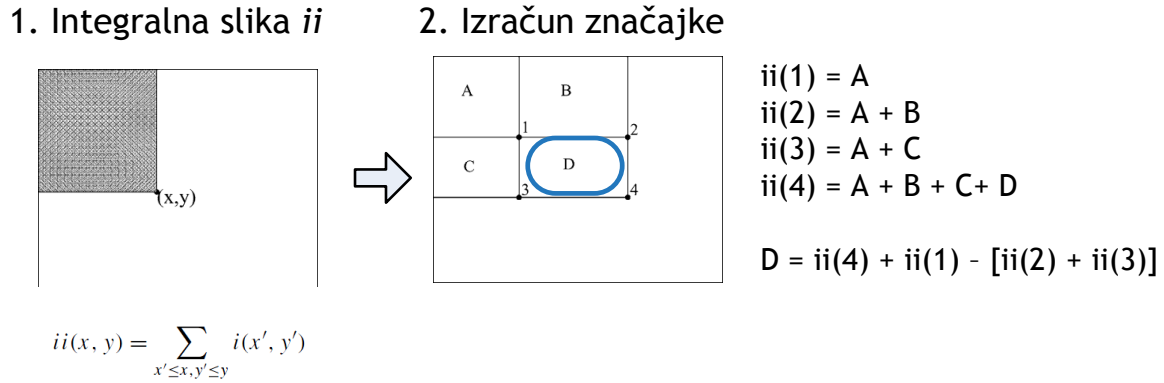
$$ii(x,y) = \sum_{x' \leq x, y' \leq y} i(x',y') \; . \tag{2.32}$$

Tako se vrijednost značajke može izračunati u samo četiri koraka prikazana na slici 2.9 desno.

Ostali radovi srodni lokalizaciji ulančavanjem uključuju [130, 138, 139]. U radu predstavljenom u [130] računalna se složenost smanjuje učenjem C ojačanih lokalizacijskih modela (engl. *boosted classifiers*) koji mogu dijeliti komponente, odnosno slabe lokalizacijske modele. Radovi predstavljeni u [138, 139] oslanjaju se na stablaste strukture, gdje čvor stabla odgovara ojačanom klasifikatoru. Ovisno o odluci u čvoru roditelju, prozor se prosljeđuje specijaliziranim lokalizacijskim modelima u čvorovima djece.



Slika 2.8: Ulančavanje osnovnih lokalizacijskih model H_i. Lokalizacijski modeli su uređeni prema složenosti, gdje lokalizacijski model H_1 odgovara najjednostavnijem modelu.



Slika 2.9: Reprezentacija integralnom slikom (lijevo). Učinkovit proračun vrijednosti značajke za slikovni regiju označenu sa D u samo četiri operacije (desno).

2.3.2 Lokalizacija učinkovitim pretraživanjem potprozora

Strategija grananja i ograničavanja (engl. *branch and bound*) [132] pronalazi prozore maksimalnog odziva na temelju svih prozora u slici. Postupak se obavlja iterativnim grananjem od većih slikovnih regija ka manjima dok se ne pronađe lokacija objekta. U svakoj se iteraciji kao kandidat za dijeljenje odabire slikovna regija s najvećom gornjom granicom odziva. Učinkovitost pretraživanja ovisi upravo o vrijednosti gornje granice odziva. Opisani pristup primjenjiv je za lokalizacijske modele čiji se odziv može razložiti kao suma odziva manjih slikovnih regija. Drugim riječima, zahtjeva se aditivnost slikovne reprezentacije.

2.3.3 Lokalizacija pretraživanjem na temelju mjere *objektnosti*

Sasvim općenito, slike mogu sadržati objekte različitih razreda. Lokalizacija na temelju mjere *objektnosti* (engl. *objectness*) [134] pripada u skupinu strategija lokalizacije gdje se prostor pretraživanja sužava predlaganjem prozora neovisno o tipu razreda za koji se provodi lokalizacija (engl. *class-independent proposals*). Navedeni pristup se temelji na mjerenju sličnosti između opisanih pravokutnika dobivenih lokalizacijom i označenih pravokutnika (od strane ljudskog agenta). Za određivanje sličnosti koriste se mjere segmentacije superpikselima, mjere *ispupčenosti* (engl. *saliency*), boje, rubova i pozicije. Model *objektnosti* uči se na podskupu primjera

za učenje koji sadrži objekte različitih razreda. Rezultantni se prozori dobivaju uzorkovanjem regija s visokim odzivom mjere *objektnosti*. Eksperimenti provedeni u okviru rada [134] pokazuju da se više od 90 posto objekata uspješno lokalizira na temelju uzorkovanja tek tisuću prozora najvećeg iznosa mjere *objektnosti*. Time se značajno smanjuje složenost pretraživanja. Prednost opisane strategije lokalizacije jest da omogućuje primjenu složenijih lokalizacijskih modela koji bi u suprotnome bili prespori ili nekompatibilni (primjerice, stroj sa potpornim vektorima uz primjenu nelinearnih jezgrenih funkcija (engl. *Support vector machine*) [140]).

2.3.4 Lokalizacija pretraživanjem na temelju segmentacije superpikselima

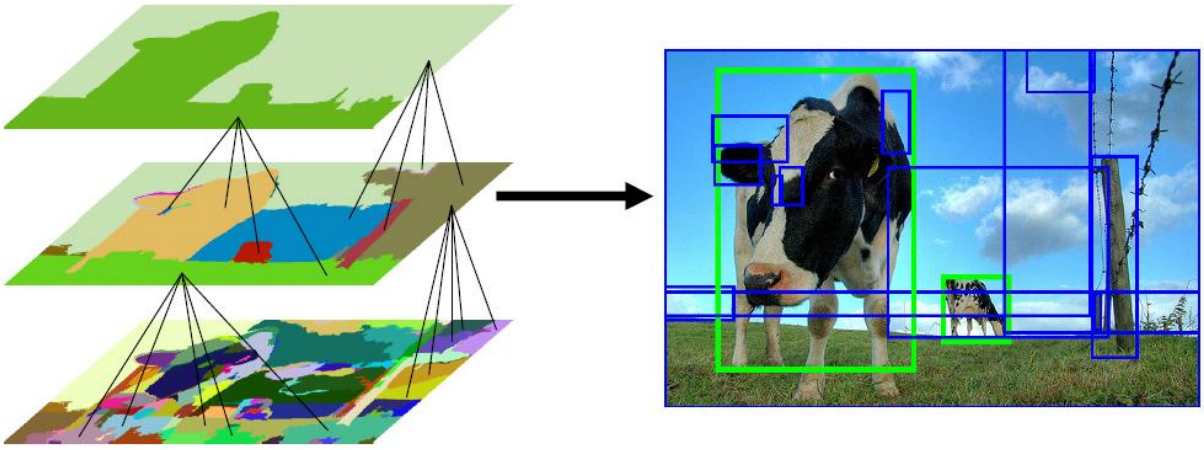
Strategija lokalizacije predložena u [44] također pripada skupini strategija neovisnih o razredu za koji se provodi lokalizacija. Slika 2.10 ilustrira postupak primjene strategije lokalizacije temeljene na segmentaciji. Sužavanje prostora pretraživanja obavlja se na temelju poligona segmentacije generiranih metodom predloženom u okviru rada [141] (poligoni dobiveni na najnižoj razini na slici 2.10). Dobiveni se poligoni koriste za izgradnju hijerarhijskog segmentacijskog stabla. Postupak izgradnje stabla uključuje iterativno „pohlepno" stapanje susjednih segmenata na temelju boja ili teksture. Rezultantni se prozori dobivaju kao opisani pravokutnici dobivenih segmentacijskih poligona.

2.3.5 Rasprava lokalizacijskih strategija

U okviru ovog odjeljka opisan je odnos lokalizacijskog pristupa ove disertacije u odnosu na strategije opisane u prethodnim odjeljcima.

Opisani pristup sličan je strategiji lokalizacije ulančavanjem predstavljene u okviru [3]. Oba pristupa primjenjuju gusto uzorkovanje značajki. Nadalje, u okviru ove disertacije također se obavlja probir slikovnih okana nalik na probir opisan u [3], gdje svaki stupanj ulančanog klasifikatora odbacuje nediskriminativne značajke. Lokalizacijski pristup ove disertacije vrši probir slikovnih okana na temelju diskriminativnih slikovnih riječi odabranih učenjem klasifikatora. Zatim se okna koja nisu dodijeljena diskriminativnim slikovnim riječima uklanjaju iz daljnjeg razmatranja. Na taj se način postiže značajno ubrzanje slično efektu prvih nekoliko razina ulančanog klasifikatora [3].

Lokalizacijske strategije koje se zasnivaju na sužavanju prostora hipoteza prijedlozima temeljenim na *objektnosti* [134] i segmentaciji [44] ostavljaju mogućnost da se objekt traženog razreda ne pojavljuje u okviru generiranih prijedloga. Time se inicijalno povećava frekvencija promašaja i prije same primjene lokalizacijskog modela za proračun odziva okna. To je poglavito važno prilikom lokalizacije malenih objekata u slabo nadziranom okruženju. Preliminarni eksperimenti na skupu podataka za prometne znakove [72] pokazali su da strategija lokaliza-



Slika 2.10: Ilustracija strategije lokalizacije na temelju segmentacije superpikselima. Preuzeto iz [44].

cije mjerom *objektnosti* ne uspjeva izolirati lokacije prometnih znakova u top 2000 prijedloga slikovnih regija [20]. S druge strane, pristup lokalizaciji „odozdo prema gore" sličan je strategijama na temelju *objektnosti* i segmentacije po broju rezultantnih regija, odnosno slikovnih okana koja potencijalno sadrže objekt od interesa. U oba slučaja, generira se nekoliko tisuća prijedloga po slici, dok se u slučaju lokalizacije „odozdo prema gore" broj potencijalnih slikovnih okana smanjuje s $8 \cdot 10^4$ na 2000 – 7000 slikovnih okana [20] koja se dalje probiru računanjem odziva rijetkog modela.

U odnosu na strategiju lokalizacije grananjem i ograničavanjem [132], lokalizacijski pristup ove disertacije također se temelji na aditivnosti reprezentacije slike u odnosu na slikovna okna. S obzirom da nelinearne normalizacije Fisherovih vektora slike narušavaju svojstvo aditivnosti slikovnih okana, u okviru ove disertacije predložena je aproksimacija prvog reda [142] odziva normalizirane slike u odnosu na Fisherov vektor okna. Primjenom predložene aproksimacije, lokalizacijski model, učen nad Fisherovim vektorima slika, može se primijeniti za proračun odziva slikovnog okna. Proračun odziva okna obavlja se putem jednostavnog skalarnog produkta kao i u slučaju bez normalizacija gdje aditivnost reprezentacije ostaje očuvana.

2.4 Slabo nadzirana lokalizacija objekata

Budući da se u okviru paradigme slabo nadzirane lokalizacije lokacije slikovnih koncepta uče na temelju oznaka na razini slike, jedan od glavnih izazova jest smanjiti računalnu složenost s obzirom na velik broj potencijalnih lokacija objekta u slici. Tako se pristupi slabo nadziranoj lokalizaciji mogu podijeliti s obzirom na strategije inicijalizacije kojima se sužava prostor pretraživanja slabo nadzirane lokalizacije na:

- postupci temeljeni na grupiranju [23]
- postupci temeljeni na prozorima odabranim prema mjeri *objektnosti* [25, 26]
- postupci temeljeni na prijedlozima segmentacije superpikselima [44] opisane u okviru

odjeljka 2.3.4: [18, 127] i pristup [24] temeljen na metodi segmentacije [143].

S druge strane, metode slabo nadzirane lokalizacije također se mogu podijeliti s obzirom na metodu optimizacije kojom se od inicijalno odabranih prozora dolazi do konačnih lokalizacijskih poligona. Tako se u literaturi mogu pronaći sljedeći pristupi:

- pristupi temeljeni na učenju na zbirkama primjeraka (engl. *Multiple Instance Learning, MIL*) [18, 22, 24, 25]
- pristupi temeljni na latentnom stroju s potpornim vektorima (engl. *Latent support vector machine, latent SVM*) [27, 128]
- pristupi temeljeni na učenju „s kraja na kraj" (engl. *end-to-end learning*) primjenom dubokih konvolucijskih mreža [28, 29, 50, 144, 145].

2.4.1 Strategije inicijalizacije slabo nadzirane lokalizacije

U okviru ovog odjeljka opisane su navedene strategije inicijalizacije koje se primjenjuju prilikom učenja lokalizacijskih modela pod slabim nadzorom.

Postupci temeljeni na grupiranju Strategija inicijalizacije temeljena na grupiranju predložena je u okviru [23]. Autori odabiru slikovne riječi koje se dominantno pojavljuju u pozitivnim slikama. Zatim se obavlja grupiranje na temelju odabranih slikovnih riječi i formiraju se opisani pravokutnici. Dobiveni opisani pravokutnici koriste se kao ulaz u algoritam učenja.

Postupci temeljeni na mjeri *objektnosti* Postupci temeljeni na mjeri *objektnosti* [120] razmatrani su u okviru radova [25, 26, 146]. Deselaers i dr. u [26] vrše probir slikovnih regija na temelju *objektnosti* te im potom dodjeljuju težine primjenom modela *objektnosti*. Model *objektnosti* uči se na odvojenom podskupu podataka koji ne sadrži primjere objekata od interesa. Za svaku sliku, inicijalno se odabire stotinjak slikovnih regija. U okviru rada predstavljenog u [25], iz svake se slike odabire po jedna slikovna regija na temelju *objektnosti* i ciljne funkcije temeljene na unutar-razrednim i među razrednim sličnostima između parova regija. Cilj optimizacije jest u pozitivnim slikama odabrati slikovne regije međusobno slične drugim regijama u pozitivnim slikama, a istovremeno što je više moguće različite u odnosu na regije prisutne u negativnim slikama. Opisana formulacija dovodi do teškog problema kombinatorne optimizacije. Iz tog se razloga, u [146] primjenjuje pojednostavljeni pristup probira slikovnih regija u kojem je naglasak na optimizaciji međurazredne sličnosti. Postupak se zasniva na maksimizaciji udaljenosti između slikovnih regija u pozitivnim slikama i najbližeg susjeda među regijama u negativnim slikama. Opisanim pojednostavljenjima zaobilazi se problem kombinatorne optimizacije. Glavni nedostatak postupaka temeljenih na *objektnosti* jest da tako odabrane slikovne regije ne sadrže nužno tražene objekte. Primjerice, na Pascal VOC2007 [131] skupu podataka, metoda [147] u 30 posto slika ne uspijeva identificirati valjane lokacije objekata.

Postupci temeljeni na segmentaciji Pristupi inicijalizaciji slabo nadziranog učenja na temelju segmentacija uključuju radove predstavljene u okviru [18, 24, 127].

Galleguillos i dr. u [24] za sužavanje prostora pretraživanja koriste postupak segmentacije predložen u [143], gdje se generiraju višestruki poligoni segmentacija na temelju kriterija stabilnosti. Dobivene se slikovne regije potom kôdiraju histogramima slikovnih riječi (engl. *Bag of Visual Words, BoVW*).

Russakovsky i dr. u [127], koriste slikovne regije dobivene postupkom selektivnog pretraživanja (engl. *Selective Search method*) [44] te ih kôdiraju metodom lokalno ograničenog linearnog kôdiranja (engl. *Locality-constrained Linear Coding, LLC*). Uz opisnike koji potencijalno sadrže objekte od interesa, istovremeno se gradi i model reprezentacije pozadine. Reprezentacija pozadine računa se nad značajkama izvan prozora generiranih postupkom selektivnog pretraživanja.

Cinbis i dr. [18] koriste slikovne regije dobivene na temelju segmentacije selektivnim pretraživanjem [44], no umjesto LLC kôdiranja koriste Fisherove vektore. Slično kao i u [127], modelira se utjecaj pozadine primjenom tzv. kontrastivnih opisnika (engl. *contrastive background descriptor*). Kontrastivni opisnici dobivaju se razlikom Fisherovih vektora pozadine (uključuje dio slike koji nije pokriven prozorima generiranim postupkom selektivnog pretraživanja) i prozora dobivenih segmentacijom. Kao i u slučaju sužavanja prostora pretraživanja na temelju mjere *objektnosti* i postupci temeljeni na poligonima dobivenim segmentacijom ne moraju nužno sadržavati tražene objekte.

2.4.2 Postupci optimizacije slabo nadzirane lokalizacije

U okviru ovog odjeljka razmatraju se postupci optimizacije prilikom slabo nadziranog učenja.

Učenje zbirkama primjeraka Učenje na temelju zbirki primjeraka (engl. *Multiple Instance Learning, MIL*) [71] uključuje pristupe slabo nadziranoj lokalizaciji predstavljene u okviru radova [18, 22, 24, 25]. Ideja učenja na temelju zbirki primjeraka jest prikazati sliku kao zbirku pojavljivanja, gdje pozitivne zbirke sadrže barem jedno pojavljivanje objekta traženog razreda, dok negativne zbirke ne sadrže niti jedno. Postupak se obavlja iterativno, gdje svaku od iteracija čine dva koraka: 1) učenje lokalizacijskog modela i 2) primjena dobivenog lokalizacijskog modela na slikovne regije u pozitivnim i negativnim slikama kako bi se dobili primjeri za učenje za narednu iteraciju.

Galleguillos i dr. u [24] primjenjuju MIL postupak učenja te na temelju BOVW histograma slikovnih regija iterativno minimiziraju funkciju cilja ojačanog klasifikatora (engl. *boosted classifier*).

U postupku predstavljenom u [25], koristi se MIL pristup, a naglasak je na činjenici da podaci imaju višemodalnu strukturu. Višemodalnost podataka, primjerice, nastaje kao posljedica

činjenice da objekti mogu biti snimljeni iz različitih perspektiva. Koristi se model elastičnih dijelova (engl. *Deformable Part Model, DPM*) [5] kao lokalizacijski model, a autori predstavljaju postupak prevencije odabira slikovnih regija u pozadini uslijed iterativnog učenja (engl. *model drift detection*).

Crowley i Zisserman u [22] koriste model linearne diskriminativne analize (engl. *Linear Discriminant Analysis, LDA*) za relokalizaciju i probir novih primjera u okviru MIL postupka, dok se kao krajnji lokalizacijski model uči model elastičnih dijelova DPM [5].

Cinbis i dr. [18] predlažu poboljšanje MIL postupka primjenom učenja s više odjeljaka (engl. *Multi-fold MIL*). Slike iz skupa za učenje dijele se u k odjeljaka, a u svakoj se iteraciji koristi $k-1$ odjeljak za učenje modela, dok preostali odjeljak služi za prikupljanje novih primjera za učenje. Opisanom optimizacijom se sprječava ponovni odabir primjera koji su već dodani u skup za učenje, čime se sprječava da MIL optimizacija zaglavi u lokalnom optimumu.

Učenje latentnim strojem s potpornim vektorima Primjeri postupaka optimizacije koji se zasnivaju na optimizaciji latentnog stroja s potpornim vektorima uključuju radove predstavljene u [27, 128].

Pandey i Lazebnik u [128] formuliraju problem slabo nadzirane lokalizacije preko modela elastičnih dijelova [5] u kojem se lokacije objekata razmatraju kao skrivene (latentne) varijable. Inicijalno, pozitivni primjeri učenja postavljaju se na veličine cjelokupnih slika te se iterativno smanjuju ka veličini opisanog pravokutnika objekta. Kako bi se postigle što bolje estimacije opisanih pravokutnika, predlaže se automatsko rezanje slikovnih regija primjenom heuristike temeljene na veličini gradijenta [148].

Bilen i dr. u [27] koriste zaglađenu verziju latentnog stroja s potpornim vektorima, tzv. *softmax* latentni *SVM*. U okviru procesa optimizacije, naglasak je na sličnosti između odabranih lokacija u slici i naučenih grupa „primjera". Grupe primjera dobivaju se algoritmom konveksnog grupiranja na temelju slika za učenje.

Učenje „s kraja na kraj" primjenom dubokih konvolucijskih mreža Radovi [50, 144, 145] temelje se na dubokim konvolucijskim mrežama i ne oslanjaju se na inicijalizacijske strategije odjeljka 2.4.1 niti na optimizaciju u vidu učenja zbirki primjeraka. Arhitekture konvolucijskih mreža primjenjivane u radovima [50, 144] učene su na ImageNet skupu podataka [93], a prilagođene su za zadatak slabo nadzirane lokalizacije pretvorbom potpuno povezanih slojeva u konvolucijske uz sažimanje najvećom [50] ili srednjom [144] vrijednošću.

Bency i dr. u [145] predlažu pristup lokalizaciji na temelju iterativne primjene potpunopovezanog sloja na poduzorkovanu konvolucijsku reprezentaciju. Dobivene slikovne regije organizirane su u stablastu strukturu, a iterativni postupak pretraživanja regija izveden je pretraživanjem zrakom (engl. *beamsearch*). Pretpostavka zrakastog pretraživanja jest da da bolje

centrirani objekti daju veće iznose odziva. Predstavljeni pristup računski je složeniji u odnosu na [50, 144], budući da zahtijeva više unaprijednih prolaza kroz mrežu.

2.4.3 Rasprava slabo nadzirane lokalizacije

Predstavljeni pristup slabo nadziranoj lokalizaciji obavlja globalno optimalno pretraživanje slikovnih okana rijetkim modelima učenim nad Fisherovim vektorima cjelokupnih slika. U tom se aspektu, razlikuje u odnosu na radove [18, 25, 26, 127] koji se zasnivaju na postupcima sužavanja prostora pretraživanja temeljenim na mjerama *objektnosti* [26] ili segmentacije [44, 143].

S obzirom na strategiju inicijalizacije slikovnih regija, slabo nadzirani pristup ove disertacije sličan je radovima [18, 128] koji također započinju učenje inicijalizacijom primjera na cjelokupnu veličinu slike, no za razliku od [18, 128], nije iterativan.

S obzirom na način reprezentacije značajki, pristup je sličan [18], gdje se za predstavljanje slikovnih regija koriste se Fisherovi vektori, no pristup [18] se zasniva na MIL optimizaciji.

Slično kao i u radovima [50, 144], pristup ove disertacije prikladan je za učenje „s kraja na kraj" ali se može primijeniti i na konvolucijske značajke izdvojene iz unaprijed naučenih arhitektura dubokih mreža kao što su VGG-E [39] i AlexNet [7]. Pristup nije iterativan te kao takav pruža mogućnost integracije s pristupima predstavljenim u [123, 128].

Poglavlje 3

Lokalizacija odozdo prema gore Fisherovim vektorima

U okviru ovog poglavlja detaljno je opisan sustav za lokalizaciju „odozdo prema gore" temeljen na Fisherovim vektorima i rijetkim klasifikacijskih modelima. Razvijeni sustav posebno je pogodan za lokalizaciju u paradigmi slabo nadziranog učenja, gdje u primjerima za učenje nisu dostupne oznake lokacija, odnosno opisani poligoni objekata (engl. *bounding polygon*). U slučaju slabo nadziranog učenja algoritam učenja najprije mora u pozitivnim slikama odrediti lokacije objekata, a zatim na temelju njih naučiti lokalizacijski model. Povoljno svojstvo reprezentacije Fisherovim vektorima je da poništava utjecaj pozadinske informacije u slici te sliku predstavlja onim značajkama koje ju čine različitom u odnosu na globalnu statistiku sadržaja slike. Shodno tome, lokalizacijski se model uči nad Fisherovim vektorima slika, a zatim se u fazi lokalizacije primjenjuje za izračun odziva slikovnih okana. Za razliku od postupaka lokalizacije čiji je pregled dan u okviru odjeljaka 2.3, pristup predložen u okviru ove disertacije obavlja globalno optimalno pretraživanje „odozdo prema gore" primjenom lokalizacijskih modela rijetkih po komponentama: od slikovnih okana na temelju kojih se određuju predikcije opisanih poligona objekata.

U nastavku poglavlja najprije je potanko opisan prvi doprinos ove disertacije, odnosno pristup slabo nadziranoj lokalizaciji temeljen na Fisherovim vektorima (odjeljak 3.1). Budući da se učinkovitost pretraživanja prostora mogućih hipoteza postiže rijetkim modelima, u okviru odjeljka 3.3.1 dan je pregled različitih funkcija regularizacije kojima se inducira rijetkost modela. Učenjem rijetkih modela identificiraju se diskriminativne Gaussove komponente (slikovne riječi) te se na temelju njih obavlja efikasan izračun odziva slikovnih okana (odjeljak 3.4.2).

Drugi i treći doprinos ove disertacije odnose se na primjenu nelinearnih normalizacija u svrhu poboljšanja lokalizacijske učinkovitosti. Najprije je dan pregled tipova nelinearnih normalizacija (odjeljak 3.2), a opisana je i metrička normalizacija po komponentama. Kako bi se omogućila efikasna primjena opisanih normalizacija u fazi lokalizacije, u okviru odjeljka 3.4.1

formalizirana je aproksimacija prvog reda koja omogućava efikasan izračun odziva slikovnog okna u slučaju normalizacija.

Četvrti doprinos ove disertacije odnosi se na reprezentacije prostornog rasporeda slikovnih okana, a razmatra se u okviru sljedećeg poglavlja.

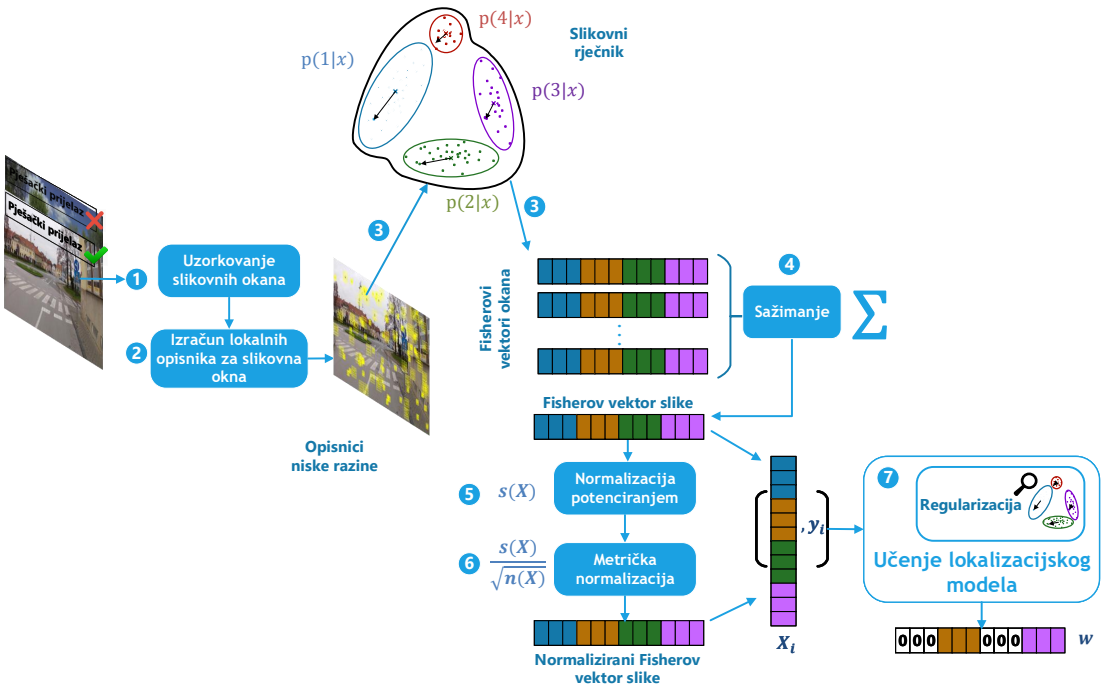
3.1 Arhitektura sustava lokalizacije odozdo prema gore

Glavni doprinos ove disertacije odnosi se na sustav slabo nadzirane lokalizacije temeljen na Fisherovim vektorima i rijetkim modelima. Postupak određivanja lokacija objekata određenog razreda podijeljen je u dvije faze. Najprije se na temelju slika i oznaka prisutnosti objekata u slikama obavlja učenje lokalizacijskog modela. Pomoću dobivenog modela obavlja se pretraživanje slikovnih okana odgovornih za prisutnost objekta u slici. Zatim se na temelju okana pozitivnog odziva formiraju se predikcije lokalizacijskih poligona.

3.1.1 Učenje lokalizacijskog modela

Shematski prikaz faze učenja sustava lokalizacije „odozdo prema gore" ilustriran je na slici 3.1 na primjeru pješačkih prijelaza kao traženog razreda objekata. Ulaz u fazu učenja čine slike iz skupa za učenje i pripadajuće oznaka prisutnosti objekta („pješački prijelaz: da / ne"). Kako bi se omogućilo učenje diskriminativnog lokalizacijskog modela, slika se predstavlja Fisherovim vektorom $\mathbf{X_i}$, a pripadajuća oznaka prisutnosti skalarom $y_i \in \{-1, 1\}$. Vrijednost $y_i = 1$ označava da slika sadrži jedan ili više objekata traženog razreda.

Postupak započinje uzorkovanjem slikovnih okana (korak 1 u slici 3.1) koja se potom predstavljaju opisnicima niske razine (korak 2). Na slici su ilustrirani SIFT opisnici, a prilikom eksperimentalnog vrednovanja koriste se još i konvolucijske značajke. Kako bi se lokalni opisnici projicirali u prostor Fisherovih vektora (korak 3) potreban je generativan model opisnika slikovnih okana. Najčešće se u tu svrhu koristi [31, 33] model raspodjele Gaussovih mješavina, GMM [34]. Model mješavine Gaussovih raspodjela posebno je pogodan budući da se uz dovoljan broj komponenata mješavine njime može aproksimirati bilo koja razdioba opisnika [107]. U perspektivi opisnika iz porodice zbirke slikovnih riječi kojoj pripadaju Fisherovi vektori, GMM odgovara slikovnom rječniku, a pojedine Gaussove komponente slikovnim riječima. Model mješavine Gaussovih raspodjela naučen je prethodno postupcima nenadziranog učenja na slučajno odabranom skupu lokalnih opisnika. Za potrebe učenja GMM-a koristi se algoritam maksimizacije očekivanja (engl. *Expectation Maximization algorithm, EM*) [98, 124]. Na slici je 3.1 prikazan pojednostavljeni 2D model mješavine Gaussovih raspodjela sa $K = 4$ komponente. U praksi dimenzionalnost odgovara dimenzionalnosti lokalnih opisnika, a koristi se i znatno veći broj slikovnih riječi. U okviru eksperimenata u poglavlju 5 koriste se modeli sa



Slika 3.1: Shematski prikaz faze učenja predstavljenog sustava lokalizacije. Kao slikovni rječnik, prikazan je model mješavine Gaussovih raspodjela. U svrhu jednostavnosti, prikazan je jednostavan 2D slučaj GMM-a sa $K = 4$ komponente, dok u stvarnosti dimenzionalnost pojedinih komponenti odgovara dimenzionalnosti lokalnih opisnika. Kao primjer lokalnih opisnika, ilustrirane su SIFT značajke koje se kôdiraju u prostor Fisherovih vektora. Blokovi Fisherovih vektora koji odgovaraju doprinosima pojedinih GMM komponenata označeni su različitim bojama. Nakon učenja rijetkog modela, vrijednosti blokova modela $\mathbf{w}$ koji odgovaraju nediskriminativnim slikovnim riječima postavljene su na ništicu.

$K = 128$ i $K = 1024$ komponente. Rezultat kôdiranja u odnosu na GMM predstavljaju Fisherovi vektori slikovnih okana, gdje je doprinos svake Gaussove komponente označen odgovarajućom bojom. U okviru četvrtog koraka na slici 3.1, obavlja se sažimanje Fisherovih vektora slikovnih okana u Fisherov vektor slike $\mathbf{X_i}$. U okviru ove disertacije, koristi se sažimanje usrednjavanjem (engl. *sum pooling*). U svrhu poboljšanja klasifikacijske i lokalizacijske točnosti, na Fisherov vektor slike mogu se primijeniti normalizacija potenciranjem (korak 5) i metrička normalizacija (korak 6).

Nakon što je obavljeno kôdiranje slika u prostor Fisherovih vektora, dobivene pozitivne ($y_i = 1$, sadrže objekt) i negativne slike ($y_i = -1$, ne sadrže objekt) koriste se kao ulaz u algoritam učenja lokalizacijskog modela (korak 7 u slici 3.1). Budući da je transformacija u prostor Fisherovih vektora kvadratna funkcija opisnika slikovnih okana (izrazi 2.24, 2.25, 2.26), Fisherovi vektori su pogodni za odvajanje linearnim diskriminativnim modelima. U okviru eksperimentalnog vrednovanja u poglavlju 5 obavlja se učenje linearnih diskriminativnih modela minimizacijom logističkog gubitka. Kako bi se osiguralo da model ispravno generalizira na skupu za testiranje, prilikom učenja primjenjuju se regularizacijske funkcije. Kako bi se smanjila računska složenost uslijed visoke dimenzionalnosti Fisherovih vektora, posebna je pažnja

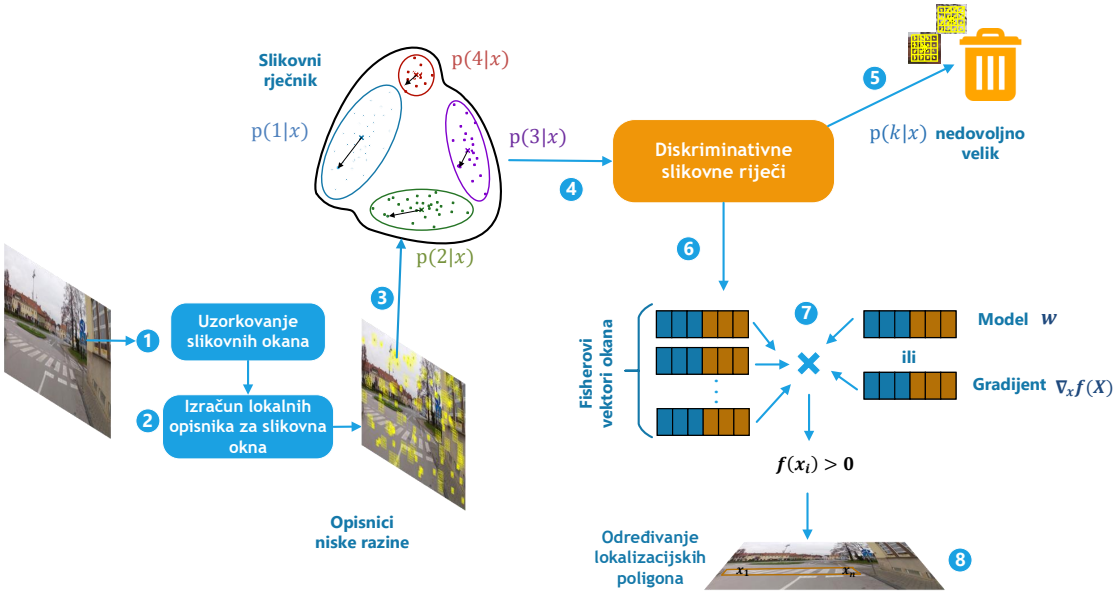
usmjerena na regularizacijske funkcije kojima se inducira rijetkost modela: ℓ_1 i $\ell_{2,1}$ regularizacije. Primjenom takvih regularizacijskih funkcija, obavlja se globalno optimalno pretraživanje diskriminativnih Gaussovih komponenata u okviru algoritma učenja [20]. Rezultat faze učenja je linearan model $\mathbf{w}$ u kojem su koeficijenti koji odgovaraju neinformativnim slikovnim riječima postavljeni na ništicu.

Linearan model $\mathbf{w}$ može se koristiti za potrebe klasifikacije slika, kada želimo pronaći sve slike koje sadrže objekte određenog razreda. Tada se odziv slike dobiva skalarnim produktom Fisherova vektora slike (koji je opcionalno normaliziran) i lokalizacijskog modela. U nastavku je opisan način primjene dobivenog modela za potrebe lokalizacije objekata.

3.1.2 Lokalizacija objekata

Slika 3.2 predstavlja shematski prikaz faze lokalizacije pristupom „odozdo prema gore". Ulaz faze lokalizacije čine slika iz skupa za testiranje i linearan model $\mathbf{w}$. Kao i u fazi učenja, postupak započinje izdvajanjem slikovnih okana (korak 1) i predstavljanjem dobivenih okana lokalnim opisnicima (korak 2). Za dobivene lokalne opisnike provodi se operacija mekog pridruživanja (engl. *soft assignment*) u odnosu na komponente modela mješavine Gaussovih raspodjela (korak 3). Na taj se način računaju vjerojatnosti $p(k|\mathbf{x})$ kojima je neki lokalni opisnik $\mathbf{x}$ dodijeljen Gaussovim komponentama indeksiranim sa $k \in [1,K]$. Analizom linearnog modela $\mathbf{w}$ utvrđene su diskriminativne Gaussove komponente značajne za odvajanje u odnosu na pozadine. Na temelju diskriminativnih komponenata, iz razmatranja se uklanjaju lokalni opisnici za koje je vjerojatnost mekog pridruživanja $p(k|\mathbf{x})$ u odnosu na odabrane komponente zanemariva (koraci 4 i 5 na slici 3.2). Na taj se način postiže značajno ubrzanje, budući da se za ta okna ne obavlja izračun Fisherovih vektora niti skalarni produkt u odnosu na model $\mathbf{w}$. Za preostale lokalne opisnike provodi se izračun Fisherovih vektora tek za maleni dio cjelokupne visokodimenzionalne reprezentacije (korak 6). Drugim riječima računaju se doprinosi prema izrazima (2.24), (2.25), (2.26) isključivo u odnosu na diskriminativne komponente. Kako bi se izračunao odziv pojedinog okna, obavlja se operacija skalarnog produkta u odnosu na lokalizacijski model $\mathbf{w}$ ili gradijent odziva normalizirane slike $\nabla_{\mathbf{x}} f(\mathbf{X})$ (korak 7 u slici 3.2). Lokalizacijski se model primjenjuje ako u fazi učenja nisu primjenjivane nelinearne normalizacije, dok se u suprotnome primjenjuje vektor gradijenta. Proračun odziva na temelju gradijenta jedan je od glavnih doprinosa ove disertacije jer omogućava efikasnu primjenu nelinearnih normalizacija u sklopu lokalizacije objekata. Na taj se način povećava točnost lokalizacije, a istovremeno se postiže i vremenska učinkovitost.

Na temelju slikovnih okana pozitivnog odziva $f(\mathbf{x}_i) > 0$ formiraju se predikcije lokalizacijskih poligona (korak 8). Na slici je takav poligon ilustriran narančastom bojom. Razmatraju se dva osnova pristupa: i) generiranje opisanih pravokutnika na temelju prostornog grafa povezanosti okana na određenom mjerilu te ii) generiranje konveksnih ljuski na temelju ujedinjenih



Slika 3.2: Shematski prikaz faze lokalizacije predstavljenog pristupa. Ulaz u fazu lokalizacije čine slika iz skupa za testiranje i lokalizacijski model **w**. Ovdje su u svrhu preglednosti uklonjene komponente modela **w** čija je vrijednost jednaka ništici. Na temelju modela **w**, određuju se diskriminativne slikovne riječi. Slikovna okna, za koje iznos vjerojatnosti pridruživanja $p(k|\mathbf{x})$ u odnosu na diskriminativne komponente nije značajan, uklanjaju se iz razmatranja.

mapa odziva preko više mjerila. Konveksne ljuske odgovaraju opisanim poligonima. U svrhu povećanja lokalizacijske točnosti, primjenjuju se algoritmi uklanjanja višestrukih lokalizacijskih poligona.

3.2 Normalizacije Fisherova vektora

Perronnin i dr. u okviru [32] predlažu poboljšanja reprezentacije Fisherovim vektorima primjenom nelinearnih normalizacija. U okviru ovog rada razmatraju se se sljedeći tipovi normalizacija:

- normalizacija potenciranjem (engl. *power normalization*) [32, 149]
- globalna metrička normalizacija (engl. *global metric normalization*) [32]
- metrička normalizacija po komponentama (engl. *intra-component normalization*) [45, 150].

U svrhu jednostavnosti notacije, u nastavku poglavlja oznakom $\mathbf{X}$ označavat će se Fisherov vektor slike, a Fisherov vektor slikovnog okna oznakom $\mathbf{x}$. Normalizacija potenciranjem provodi se zasebno za svaki element Fisherova vektora X_d, gdje $d \in [1 \ldots K(2D+1)]$, pri čemu K označava broj GMM komponenata (slikovnih riječi), a D dimenzionalnost lokalnih opisnika:

$$s(X_d) = \text{sign}(X_d)|X_d|^{\rho}, \text{ gdje je } 0 < \rho < 1\,. \tag{3.1}$$

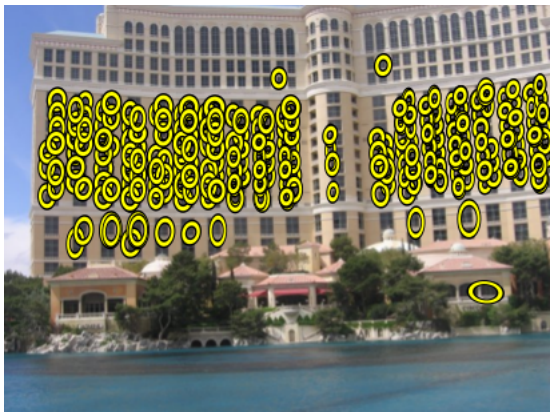
Obično se za iznos parametra ρ odabire vrijednost 0.5, tako da gornja operacija odgovara predznačenom korijenovanju (engl. *signed square routing*). Uloga normalizacije potenciranjem je višestruka. Jedna od prednosti odnosi se na primjenu pretpostavke da su lokalni opisnici $\{\mathbf{x_t},\, t = 1 \ldots T\}$ nezavisni i jednoliko raspodijeljeni (engl. *independent and identically distributed, i.i.d.*). U stvarnosti to često nije slučaj, posebice ako se opisnici niže razine preklapaju [151]. Perronin i dr. u okviru [152] pokazuju da se normalizacijom potenciranjem umanjuje utjecaj opisnika dodijeljenih karakterističnim slikovnim riječima, a koja se natprosječno često pojavljuju u slici (engl. *bursty visual features*). Takve komponente imaju veći doprinos potpisu slike. Normalizacijom potenciranjem umanjuje se doprinos takvih elemenata X_d, gdje $s(X_d) < X_d$. Time se umanjuje utjecaj potencijalno neispravne i.i.d. pretpostavke slikovnih okana. Navedena normalizacija također se može promatrati u kontekstu jezgrene funkcije $K(\mathbf{X},\mathbf{X}') = \langle \mathbf{s}(\mathbf{X}), \mathbf{s}(\mathbf{X}') \rangle$. S porastom broja Gaussovih komponenata Fisherovi vektori postaju rjeđi. Taj efekt posebice je značajan kada se Fisherov vektor računa za manju slikovnu regiju, gdje manji broj lokalnih opisnika biva dodijeljen pojedinim Gaussovim komponentama sa značajnom vjerojatnošću $p(k|\mathbf{x})$. U slučajevima gdje su vrijednosti tih vjerojatnosti bliske ništici $p(k|\mathbf{x}) \approx 0$, odgovarajući doprinosi tih komponenti Fisherovu vektoru izrazito su maleni. Primjenom normalizacije potenciranjem povećava vrijednost takvih elemenata X_d, gdje $s(X_d) > X_d$. Na taj način normalizacija potenciranjem smanjuje stupanj rijetkosti Fisherova vektora $\mathbf{X}$ te ga čini pogodnijim za uspoređivanje skalarnim produktom u okviru jezgrene funkcije.

Nakon normalizacije potenciranjem, na vektor $\mathbf{s}(\mathbf{X})$ primjenjuje se metrička normalizacija. Globalna metrička normalizacija projicira Fisherov vektor na jediničnu sferu dijeleći ga s $\sqrt{n(\mathbf{X})}$, gdje je izraz $n(\mathbf{X})$ definiran s:

$$n(\mathbf{X}) = s(\mathbf{X})^T s(\mathbf{X}) = \sum_{d=1}^{K(2D+1)} s(X_d)^2 \,. \tag{3.2}$$

Ovdje je opisana ℓ_2 norma, no metrička se normalizacija u načelu može provoditi i dijeljenjem s ℓ_p normom [33]. Autori u [33] pokazuju da reprezentacija Fisherovim vektorima poništava utjecaj pozadinskog sadržaja u slici. S druge strane, slike mogu imati različit udio sadržaja vezanog uz razred traženog objekta. Tako, primjerice, slike koje sadrže isti objekt različitih veličina (mjerila) neće imati jednaki Fisherov potpis. Primjenom metričke normalizacije umanjuju se razlike među reprezentacijama slika koje sadrže različite veličine objekata [32].

Uz globalnu metričku normalizaciju u okviru ovog rada razmatrana je i metrička normalizacija na razini slikovne riječi, odnosno Gaussove komponente [45]. Neka $\mathbf{X}^k \in \mathbb{R}^{2D+1}$ označava dio Fisherova vektora koji obuhvaća gradijente (2.24), (2.25) i (2.26) u odnosu na k-tu komponentu. Metrička normalizacija po komponentama provodi se dijeljenjem doprinosa svake komponente $\mathbf{X}^k$ s odgovarajućom normom $\sqrt{n(\mathbf{X}^k)}$. Definicija norme po komponentama ekvi-



Slika 3.3: Efekt eksplozije slikovne riječi (engl. *burstiness effect*): ilustrirani su lokalni opisnici dodijeljeni dominantnoj slikovnoj riječi. Preuzeto iz [46].

valentna je definiciji globalne norme (3.2):

$$n(\mathbf{X}^k) = \sum_{d=(k-1)(2D+1)}^{k(2D+1)} s(X_d)^2 \,. \tag{3.3}$$

Po komponentama ℓ_2-normalizirani Fisherov vektor tada odgovara:

$$\frac{1}{\sqrt{K}} \cdot \left[\frac{\mathbf{s}(\mathbf{X}^1)}{n(\mathbf{X}^1)} \cdots \frac{\mathbf{s}(\mathbf{X}^k)}{n(\mathbf{X}^k)} \cdots \frac{\mathbf{s}(\mathbf{X}^K)}{n(\mathbf{X}^K)} \right] . \tag{3.4}$$

Množenje faktorom $1/\sqrt{K}$ osigurava jediničnu normu cjelokupnog Fisherova vektora uz uvjet da je barem jedno slikovno okno dodijeljeno svakoj slikovnoj riječi. Uloga normalizacije po komponentama je da umanji potencijalni utjecaj slučajeva gdje pojedina komponenta k daje natprosječno velik doprinos Fisherovu vektoru i kao takva dominira nad skalarnim produktom s modelom $\mathbf{w}$ ili nekih drugim Fisherovim vektorom [45]. U literaturi se takav slučaj naziva „efektom eksplozije slikovne riječi" (engl. *burstiness effect*) [46]. Slika 3.3 ilustrira slučaj gdje slikovna riječ, odgovorna za generiranje slikovnih okana na prozorima, ima natprosječno visok doprinos Fisherovu vektoru slike. Normalizacija potenciranjem također nastoji umanjiti utjecaj tog efekta, smanjujući vrijednosti elemenata Fisherova vektora koje imaju natprosječno velik doprinos. Rezultat potenciranja jest Fisherov vektor u kojem je doprinos takvih komponenata umanjen, no još uvijek je veći u odnosu na ostale komponente slikovnog rječnika te može dominirati prilikom proračuna odziva slike u odnosu na klasifikacijski model. Metričkom normalizacijom po komponentama doprinosi svih komponenata svode se na jediničnu normu čime se negativan utjecaj takvih komponenti na klasifikacijski odziv posve uklanja [45].

3.3 Lokalizacija rijetkim diskriminativnim modelima

U fazi lokalizacije obavlja se gusto uzorkovanje slikovnih okana i kôdiranje u prostor visokodimenzionalnih Fisherovih vektora. Iz navedenog je razloga, glavni izazov postići računski efikasan proračun odziva slikovnih okana ne bi li se identificirala okna odgovorna za prisutnost objekta u slici. Kako bi se omogućio efikasan proračun odziva okna, koriste se rijetki modeli. Prednosti rijetkih modela navedene su u nastavku:

- U okviru procesa učenja rijetkih modela obavlja se probir značajki.
  Pretpostavka je da su specifične slikovne riječi odgovorne za generiranje slikovnih okana na objektima, odnosno pripadajućih lokalnih opisnika. Učenjem rijetkih modela obavlja se globalno optimalno pretraživanje slikovnih riječi kako bi se identificirale riječi zaslužne za generiranje okana koja ukazuju na prisutnost objekta u slici.
- Rijetki modeli rezultiraju manjom računalnom složenosti izvođenja.
  Fisherov vektor dobiva se konkaterniranjem gradijenata (2.24), (2.25), (2.26) u odnosu na određenu Gaussovu komponentu. Doprinosi pojedinih komponenti međusobno su nezavisni, a shodno tome i odgovarajuće komponente modela $\mathbf{w}$. Kako se učenjem rijetkih modela smanjuje efektivna dimenzionalnost (koeficijenti koji odgovaraju neinformativnim slikovnim riječima imaju vrijednost ništice), prilikom proračuna Fisherova vektora slikovnog okna računaju se doprinosi isključivo u odnosu na diskriminativne komponente. Na taj se način smanjuje i složenost skalarnog produkta u odnosu na Fisherov vektor okna.

Prilikom učenja rijetkog modela $\mathbf{w}$, primjenjuje se neka od regularizacijskih funkcija koje induciraju rjetkoću modela. Sasvim općenito, uloga regularizacijske funkcije jest sprečavanje prenaučenosti (engl. *overfitting*) [34]. Koncept prenaučenosti odnosi se na slučaj gdje model klasificira podatke za učenje bez pogreške, no na skupu za testiranje daje lošiju performansu. Veliki kapacitet modela dovodi do velike varijance algoritma učenja, uslijed čega se algoritam prilagođava šumu u podacima. Šum u podacima najčešće je posljedica nepreciznosti, pogrešaka ili subjektivnosti u označavanju podatka. Regularizacijom se smanjuju vrijednosti koeficijenata modela $\mathbf{w}$, odnosno pojednostavljuje se model kako bi bolje generalizirao na još neviđenim podacima.

Neka $\mathbf{X_i}$ označava i-ti primjer za učenje, a y_i njegovu oznaku pripadnosti razredu objekata ili pozadina za slučaj binarne klasifikacije. Oznakom $J(\mathbf{w}, \mathbf{X_i}, y_i)$ obilježavamo funkciju cijene (engl. *cost function*) koju prilikom učenja modela $\mathbf{w}$ minimiziramo: $\min_{\mathbf{w}} J(\mathbf{w}, \mathbf{X_i}, y_i)$. Kako bi se smanjio utjecaj prenaučenosti, prilikom minimizacije funkcije cijene $J(\mathbf{w}, \mathbf{X_i}, y_i)$ pridodaje se regularizacijska funkcija u obliku $\lambda \cdot \mathscr{R}(\mathbf{w})$:

$$\ell(\mathbf{w}, \mathbf{X}, \mathbf{y}) = \sum_i J(\mathbf{w}, \mathbf{X_i}, y_i) + \lambda \cdot \mathscr{R}(\mathbf{w}) \,. \tag{3.5}$$

Parametar λ u gornjem izrazu označava regularizacijski faktor koji utječe na ravnotežu između dva cilja učenja: i) da model dobro nauči decizijsku hiperravninu između razreda te ii) da dobiveni model ne bude previše složen, odnosno da se umanji utjecaj prenaučenosti. Vrijednost parametra λ najčešće se određuje unakrsnom provjerom (engl. *cross-validation*) [99]. U praksi se najčešće provodi *n*-terostruka unakrsna provjera (engl. *n-fold cross validation*). Skup podataka za učenje podijeli se u *n* odjeljaka. Za svaku vrijednost parametra λ računa se prosječna klasifikacijska performansa preko svih *n* odjeljaka. Proces se obavlja iterativno tako da se u svakoj od *n* iteracija za učenje koristi $n-1$ odjeljak, a testiranje se provodi na jednom odjeljku. Vrijednost parametra koja odgovara najboljoj klasifikacijskoj performansi koristi se prilikom učenja krajnjeg modela.

3.3.1 Tipovi regularizacijskih funkcija

U okviru ovog rada razmatraju se sljedeći tipovi regularizacijskih funkcija:

- ℓ_2 regularizacija (za usporedbu s rijetkim regularizacijskim funkcijama)
- ℓ_1 regularizacija
- $\ell_{2,1}$ regularizacija po komponentama (blokovima) koja uključuje ℓ_2 regularizaciju unutar komponente te ℓ_1 između komponenti.

ℓ_2 regularizacija (engl. *Tikhonov regularization, weight decay*) [34] definirana je sljedećim izrazom:

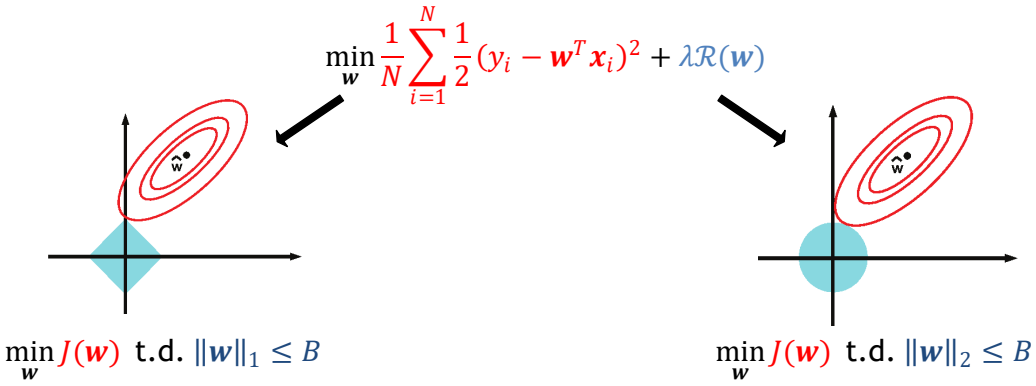
$$\ell_2(\mathbf{w}) = \|\mathbf{w}\|_2^2 = \sum_i w_i^2 = \mathbf{w}^T\mathbf{w} \,. \tag{3.6}$$

Minimizacijom izraza (3.6), kažnjava se suma veličina koeficijenata vektora $\mathbf{w}$ te se time dobiva model s manjim iznosima koeficijenata.

ℓ_1 regularizacija (engl. *least absolute shrinkage and selection operator, lasso*) definirana je na sljedeći način:

$$\ell_1(\mathbf{w}) = \|\mathbf{w}\|_1 = \sum_i |w_i| \,. \tag{3.7}$$

Slika 3.4 [34] ilustrira razlike između ℓ_1 i ℓ_2 tipova regularizacije za problem minimizacije kvadrata odstupanja kao funkcije gubitka $J(\mathbf{w})$. Problem minimizacije regularizirane funkcije gubitka $\min_w J(\mathbf{w}) + \lambda\|\mathbf{w}\|_1$ može se zapisati u obliku optimizacijske funkcije s linearnim ograničenjima $\min_w J(\mathbf{w})$ t. d. $\|w\|_1 \leq B$, gdje B označava gornju granicu na vrijednost koeficijenata. Analogno, za ℓ_2 regularizaciju, optimizacijski problem odgovara $\min_w J(\mathbf{w})$ t. d. $\|\mathbf{w}\|_2 \leq B$. Optimalno rješenje navedenih problema dobiva se u prvoj točki gdje se plohe funkcija gubitka i funkcija ograničenja dodiruju. U slučaju ℓ_1 regularizacije, intuitivno je jasno da će kutovi plohe ograničenja biti točke presjecišta jer su najistureniji. Kutovi plohe ograničenja odgovaraju rješenjima na koordinatnim osima, gdje neka od značajki poprima vrijednost ništice. U slučaju ℓ_2 regularizacije, ploha ograničenja odgovara kružnici koja može dodirnuti plohu optimizacijske



Slika 3.4: Usporedba ℓ_1 regularizacije (lijevo) i ℓ_2 regularizacije (desno) za problem najmanjih kvadrata (engl. *Least Squares, LS*) $J(\mathbf{w}) = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{2}(y_i - \mathbf{w}^T \cdot \mathbf{x}_i)^2$. Prilagođeno prema [34].

funkcije u bilo kojoj točki. Drugim riječima, rijetka rješenja nemaju prednost kao u slučaju ℓ_1 regularizacije.

Rijetka $\ell_{2,1}$ regularizacija [34, 36] podrazumijeva specifičnu strukturu podataka, gdje su podaci podijeljeni u nezavisne blokove. Iz tog razloga, posebno je pogodna za učenje modela nad Fisherovim vektorima jer omogućava izravan probir diskriminativnih slikovnih riječi. Neka $\mathbf{w}^k$ označava blok modela $\mathbf{w}$ koji odgovara komponenti indeksiranoj s k, $k \in [1 \ldots K]$, gdje K označava broj slikovnih riječi. Tada je $\ell_{2,1}$ regularizacija definirana na sljedeći način:

$$\ell_{2,1}(\mathbf{w}) = \sum_{k=1}^{K} \|\mathbf{w}^k\|_2 \,. \tag{3.8}$$

Primjenom ℓ_2 regularizacije unutar doprinosa pojedine komponente $\mathbf{w}^k$, postiže se da klasifikator po toj komponenti dobro generalizira nad podacima za testiranje, odnosno da korektno nauči granice između razreda. S druge strane, primjenom ℓ_1 regularizacije između komponenti postiže se efekt da se svi koeficijenti neke nediskriminativne komponente $\mathbf{w}^k$ postave u ništicu.

3.4 Efikasan proračun odziva slikovnih okana

U okviru ovog odjeljka razmatra se problem efikasnog proračuna odziva slikovnih okana primjenom rijetkog modela. Analogno prethodnim odjeljcima, koristi se notacija $\mathbf{X}$ za označavanje Fisherova vektora slike i $\mathbf{x}$ za označavanje Fisherova vektora slikovnog okna. Jednostavnosti radi, razmatra se slučaj binarne klasifikacije s dva razreda, gdje $f(\mathbf{x})$ označava odziv okna, dok $f(\mathbf{x}) > 0$ označava da okno ukazuje na prisutnost traženog razreda objekata u slici. Pristup je primjenjiv i za problem kada su u slici prisutni objekti različitih razreda uz supstituciju $f(\mathbf{x}) \rightarrow f(\mathbf{x}|c)$, gdje c označava pojedini razred.

Za efikasan proračun odziva slikovnog okna razlikuju se dva osnovna slučaja: i) model je učen nad nenormaliziranim Fisherovim vektorima slika te ii) model je učen nad normaliziranim

Fisherovim vektorima kao što je opisano u odjeljku 3.2. U slučaju bez normalizacija, proračun odziva je jednostavan jer vrijedi svojstvo aditivnosti opisano jednadžbom* (2.27). Klasifikacijski odziv slike odgovara izrazu:

$$f_{\text{lin}}(\mathbf{X}) = \mathbf{w}^T \cdot \mathbf{X} \,. \tag{3.9}$$

Uz primjenu svojstva aditivnosti (2.27), klasifikacijski odziv slike može se prikazati kao suma odziva okana $\mathbf{x_i}$, odnosno $f(\mathbf{X}) = \mathbf{w}^T \cdot \mathbf{X} = \sum_i \mathbf{w}^T \cdot \mathbf{x_i}$. Odziv pojedinačnog okna $\mathbf{x}$ u tom slučaju odgovara jednostavnom skalarnom produktu s modelom $\mathbf{w}$:

$$f_{\text{lin}}(\mathbf{x}) = \mathbf{w}^T \cdot \mathbf{x} \,. \tag{3.10}$$

U slučaju normalizacija, situacija je nešto složenija. Radi jednostavnosti notacije, u nastavku odjeljka podrazumijeva se korištenje globalne metričke normalizacije (3.2), no analogno razmatranje vrijedi i za metričku normalizaciju po komponentama (3.3). Klasifikacijski odziv slike u tom slučaju odgovara:

$$f_{\text{norm}}(\mathbf{X}) = \mathbf{w}^T \cdot \frac{\mathbf{s}(\mathbf{X})}{\sqrt{n(\mathbf{X})}} \,. \tag{3.11}$$

Nelinearne normalizacije poništavaju svojstvo aditivnosti, odnosno $\mathbf{w}^T \cdot \mathbf{s}(\mathbf{X}) / \sqrt{\mathbf{n}(\mathbf{X})} \neq \sum_i \mathbf{w}^T \cdot \mathbf{s}(\mathbf{x_i}) / \sqrt{\mathbf{n}(\mathbf{x_i})}$. Jedno moguće rješenje jest izračunati doprinos okna izravno:

$$f_{\text{norm}}(\mathbf{x}) = f_{\text{norm}}(\mathbf{X}) - f_{\text{norm}}(\mathbf{X} - \mathbf{x}) \,. \tag{3.12}$$

Međutim, takav bi proračun odziva okna uključivao sljedeće operacije za svako okno u slici: i) oduzimanje Fisherova vektora okna u odnosu na Fisherov vektor slike $\mathbf{X} - \mathbf{x}$, ii) primjenu normalizacije potenciranjem $s(\mathbf{X} - \mathbf{x})$, ii) primjernu metričke normalizacije $s(\mathbf{X} - \mathbf{x}) / \sqrt{n(\mathbf{X} - \mathbf{x})}$ te iv) skalarni produkt s modelom $\mathbf{w}^T \cdot s(\mathbf{X} - \mathbf{x}) / \sqrt{n(\mathbf{X} - \mathbf{x})}$. Međutim, s obzirom na potencijalno velik broj okana u slici, nije efikasno provoditi skupe operacije korijenovanja i množenja na razini pojedinačnog okna. Zbog toga se kao treći doprinos ove disertacije predlaže aproksimacija prvog reda koja omogućava efikasan proračun odziva okna u slučaju s normalizacijama.

3.4.1 Gradijent odziva normalizirane slike

U okviru ovog odjeljka razmatra se aproksimacija odziva slikovnog okna u slučaju kada je model učen na normaliziranim Fisherovim vektorima slika. Umjesto skupih i složenih operacija koje uključuje pristup opisan jednadžbom (3.12), odziv slikovnog okna može se prikazati

*Uz supstituciju notacije $\Phi_{\boldsymbol{\theta}}(\mathbf{X}) \rightarrow \mathbf{X}$, $\phi_{\boldsymbol{\theta}}(\mathbf{x_i}) \rightarrow \mathbf{x_i}$.

aproksimacijom prvog reda na temelju razvoja funkcije f_{norm} u Taylorov red oko $\mathbf{X}$ [142]:

$$f_{\text{norm}}(\mathbf{X}+\mathbf{x}) \approx f_{\text{norm}}(\mathbf{X}) + \nabla_{\mathbf{x}} f_{\text{norm}}(\mathbf{X})^T \cdot \mathbf{x} \,. \tag{3.13}$$

Izraz $\nabla_{\mathbf{x}} f_{\text{norm}}(\mathbf{X})$ označava gradijent odziva normalizirane slike $f_{\text{norm}}(\mathbf{X})$ u odnosu na nenormalizirani Fisherov vektor okna $\mathbf{x}$. Shodno izrazu (3.13), doprinos okna $\mathbf{x}$ može se tada prikazati kao:

$$f_{\text{grad}}(\mathbf{x}) \approx f_{\text{norm}}(\mathbf{X}+\mathbf{x}) - f_{\text{norm}}(\mathbf{X}) \approx \nabla_{\mathbf{x}} f_{\text{norm}}(\mathbf{X})^T \cdot \mathbf{x} \,. \tag{3.14}$$

U nastavku je dan izvod vektora gradijenta za pojedinu dimenziju d, $\nabla_{x_d} f_{\text{norm}}(\mathbf{X}) = \partial f_{\text{norm}}(\mathbf{X}) / \partial x_d$.

$$\frac{\partial f_{\text{norm}}(\mathbf{X})}{\partial x_d} = \frac{\partial f_{\text{norm}}(\mathbf{X})}{\partial \mathbf{X}} \cdot \frac{\partial \mathbf{X}}{\partial x_d} \,. \tag{3.15}$$

Derivacija nenormaliziranog Fisherova vektora slike u odnosu na d-ti element Fisherova vektora slikovnog okna $\partial \mathbf{X} / \partial x_d$ odgovara vektoru u kojem su svi elementi ništice, osim elementa na d-toj poziciji koji je jednak jedinici. Gradijent se, stoga, može zapisati kao:

$$\frac{\partial f_{\text{norm}}(\mathbf{X})}{\partial x_d} = \frac{\partial f_{\text{norm}}(\mathbf{X})}{\partial X_d} \,. \tag{3.16}$$

Na temelju izraza (3.11), za izvod derivacije $\partial f_{\text{norm}}(\mathbf{X}) / \partial X_d$, potrebno je izračunati derivacije normalizacija potenciranjem i metričke normalizacije. Derivacija normalizacije potenciranjem $\mathbf{s}(\mathbf{X})$ u odnosu na element Fisherova vektora slike X_d dana je izrazom:

$$\frac{\partial \mathbf{s}(\mathbf{X})}{\partial X_d} = [0 \ldots \rho |X_d|^{\rho-1} \ldots 0] \,. \tag{3.17}$$

Derivacija globalne metričke normalizacije $n(\mathbf{X}) = s(\mathbf{X})^T s(\mathbf{X})$ dana je izrazom:

$$\frac{\partial n(\mathbf{X})}{\partial X_d} = 2\mathbf{s}(\mathbf{X}) \frac{\partial \mathbf{s}(\mathbf{X})}{\partial X_d} = 2s(X_d)\, \rho |X_d|^{\rho-1} \,. \tag{3.18}$$

Izvod derivacije za metričku normalizaciju po komponentama (3.3) analogan je gornjem izrazu za globalnu metričku normalizaciju, gdje se izraz $n(\mathbf{X})$ supstituira sa $n(\mathbf{X}^k)$, a izraz $\mathbf{s}(\mathbf{X})$ sa $\mathbf{s}(\mathbf{X})^{\mathbf{k}}$, gdje $\mathbf{X}^k$ odgovara bloku Fisherova vektora dobivenog u odnosu na k-tu komponentu modela razdiobe Gaussovih mješavina. Opisana supstitucija je moguća jer su doprinosi u odnosu na pojedine Gaussove komponente međusobno nezavisni. Derivacija izraza (3.16) može

se prikazati kao:

$$\begin{aligned}\frac{\partial f(\mathbf{X})}{\partial X_d} &= \frac{\partial \mathbf{w}^\top \mathbf{s}(\mathbf{X})/\sqrt{n(\mathbf{X})}}{\partial X_d} \\ &= \frac{1}{\sqrt{n(\mathbf{X})}} \frac{\partial \sum_d w_d \cdot s(X_d)}{\partial X_d} + \mathbf{w}^\top \mathbf{s}(\mathbf{X}) \frac{\partial [n(\mathbf{X})]^{-0.5}}{\partial X_d} \\ &= \frac{w_d}{\sqrt{n(\mathbf{X})}} \frac{\partial s(X_d)}{\partial X_d} - 0.5 \cdot \frac{\mathbf{w}^\top \mathbf{s}(\mathbf{X})}{[n(\mathbf{X})]^{1.5}} \frac{\partial n(\mathbf{X})}{\partial X_d}\end{aligned} \tag{3.19}$$

Primjenom izraza (3.17) i (3.18) dobiva se:

$$\frac{\partial f(\mathbf{X})}{\partial X_d} = \frac{w_d}{\sqrt{n(\mathbf{X})}} \rho |X_d|^{\rho-1} - 0.5 \cdot \frac{\mathbf{w}^\top \mathbf{s}(\mathbf{X})}{[n(\mathbf{X})]^{1.5}} \cdot 2 \cdot s(X_d) \cdot \rho |X_d|^{\rho-1} \tag{3.20}$$

$$= \frac{\rho |X_d|^{\rho-1}}{\sqrt{n(\mathbf{X})}} \left(w_d - \frac{\mathbf{w}^\top \mathbf{s}(\mathbf{X})}{n(\mathbf{X})} \cdot s(X_d) \right) \tag{3.21}$$

$$= \frac{\rho |X_d|^{\rho-1}}{\sqrt{n(\mathbf{X})}} \left(w_d - \frac{f(\mathbf{X})\, s(X_d)}{\sqrt{n(\mathbf{X})}} \right) \tag{3.22}$$

Valja napomenuti da je gornji izraz nedefiniran u slučaju kada je $X_d = 0$ (uz standardnu vrijednost $\rho = 0.5$, dobiva se $|X_d|^{\rho-1} = 1/\sqrt{|X_d|}$). To je iznimno rijedak slučaj budući da su Fisherovi vektori slika gusti, odnosno dobiveni sažimanjem operacijom usrednjavanja (engl. *sum pooling*) Fisherovih vektora pojedinih okana. Ipak, kako bi se postigla numerička stabilnost postupka u slučaju $X_d = 0$, vrijednost vektora gradijenta na poziciji d postavlja se na vrijednost ništice.

U slučaju metričke normalizacije po komponentama, klasifikacijski odziv slike može se prikazati kao suma odziva po komponentama, odnosno $f_{\text{norm}}(\mathbf{X}) = \sum_k f_{\text{norm}}(\mathbf{X}^k)$. Budući da odzivi po komponentama $f_{\text{norm}}(\mathbf{X}^k)$ imaju jednak oblik kao i odziv cjelokupne slike, gradijent se može proračunati na analogan način za svaku komponentu.

3.4.2 Optimizacije izračuna odziva okna

U okviru ovog odjeljka razmatraju se optimizacije cjelokupnog postupka lokalizacije počevši od izračuna Fisherovih vektora slikovnih okana do izračuna njihovih odziva. Nadalje, podrazumijeva se primjena regularizacijskih funkcija koje induciraju rjetkoću modela prilikom učenja lokalizacijskog modela. Rezultantni lokalizacijski model $\mathbf{w}$ odlikuje visok stupanj rijetkosti.

Optimizacije su podijeljene u tri koraka. U prvom koraku obavlja se analiza rijetkog modela $\mathbf{w}$. Analogno Fisherovim vektorima nad kojima je obavljeno učenje modela, model se može prikazati kao vektor blokova $\mathbf{w}^k$, $k \in [1 \ldots K]$, gdje $\mathbf{w}^k$ označava dio modela koji odgovara k-toj komponenti modela raspodjele Gaussovih mješavina. Analizom modela utvrđuju se blokovi $\mathbf{w}^k$ s normom većom od ništice $n(\mathbf{w}^k) > 0$. Dobiveni blokovi odgovaraju diskrimina-

tivnim slikovnim riječima, odnosno komponentama modela mješavine Gaussovih raspodjela. Skup diskriminativnih slikovnih riječi obilježava se oznakom $\mathbf{S} = \{k\}$ t. d. $n(\mathbf{w}^k) > 0$, a broj diskriminativnih slikovnih riječi (kardinalnost skupa $|\mathbf{S}|$) izrazom K_w.

U drugom koraku, za sva slikovna okna u slici, obavlja se proračun vjerojatnosti $p(k|\mathbf{x})$ (2.23) kojom je lokalni opisnik pojedinog slikovnog okna $\mathbf{x}$ dodijeljen GMM komponenti indeksiranoj sa k. Slikovna okna $\mathbf{x}$, za koja je vjerojatnost pridruživanja $p(k|\mathbf{x})$ u odnosu na diskriminativne slikovne komponente $k \in \mathbf{S}$ zanemarivo mala, uklanjaju se iz daljnjeg razmatranja i za njih se ne računaju Fisherovi vektori niti odzivi modela. U praksi se koristi uvjet $p(k|\mathbf{x}) > 1/K$, gdje K označava ukupan broj slikovnih riječi. Ovisno o količini informacije vezane uz objekte od interesa u slici, ovom se optimizacijom može postići znatno ubrzanje.

U trećem se koraku razmatraju slikovna okna koja zadovoljavaju uvjet $p(k|\mathbf{x}) > 1/K$. Za odabrana okna, evaluiraju se isključivo elementi Fisherova vektora koji odgovaraju diskriminativnim komponentama, $\{\phi_{\alpha_k}(\mathbf{x}), \phi_{\boldsymbol{\mu}_k}(\mathbf{x}), \phi_{\boldsymbol{\sigma}_k}(\mathbf{x})\}, k \in \mathbf{S}$ prema jednadžbama 2.24), (2.25) i (2.26). Ostalih $K - K_\mathrm{w}$ neinformativnih komponenti uklanja se iz razmatranja te se time postiže dodatno ubrzanje za faktor K/K_w.

Predloženi niz optimizacija primjenjiv je u slučaju kada se odziv slikovnog okna računa skalarnim produktom u odnosu na i) model $\mathbf{w}$ prema izrazu (3.10), odnosno ii) gradijent odziva normalizirane slike $\nabla_\mathbf{x} f_\mathrm{norm}(\mathbf{X})$ prema (3.14). Optimizacije su također primjenjive u slučaju izravnog proračuna odziva $f_\mathrm{norm}(\mathbf{x})$ preko (3.12) ako se koristi metrička normalizacija po komponentama (3.3), (3.4). Doprinosi pojedinih komponenti u Fisherovu vektoru međusobno su nezavisni i svaka se komponenta $\mathbf{X}^k$ dijeli pripadajućom normom $\sqrt{n(\mathbf{X}^k)}$. Drugim riječima, nediskriminativne komponente $k \notin \mathbf{S}$ nemaju utjecaj na odziv Fisherova vektora pojedinog okna. U slučaju izravnog proračuna (3.12) i globalne metričke normalizacije (3.2), predložene optimizacije se ne primjenjuju budući da i nediskriminativne komponente utječu na vrijednost norme Fisherova vektora slike $n(\mathbf{X})$.

3.5 Određivanje lokalizacijskih poligona

Poligoni lokalizacije određuju se na temelju okana pozitivnog odziva na dva osnovna načina:

- pristupom temeljenim na pojedinačnom mjerilu
- pristupom temeljenim na ujedinjenim mapama odziva preko više mjerila.

3.5.1 Određivanje lokalizacijskih poligona na temelju pojedinačnih mjerila

Postupak određivanja lokalizacijskih poligona na temelju pojedinačnih mjerila prikazan je u okviru algoritma 1. Pristup odvojeno razmatra okna uzorkovana na pojedinačnim mjerilima $\mathbf{M}$.

Najprije se iz skupa svih okana **S** izdvaja T okana najvećeg odziva koja se pohranjuju u skup **D** (redak 1). Zatim se za svako od mjerila m iz skupa **M** izdvajaju slikovna okana uzorkovana na tom mjerilu u skup $\mathbf{D}_m$. Na temelju odabranih okana, konstruira se prostorni graf povezanosti G, gdje su dva okna uzorkovana na istom mjerilu povezana ako je njihov presjek veći u odnosu na prag P definiran kao parametar algoritma (redak 5). U praksi se P postavlja na određeni postotak veličine okna na promatranom mjerilu, primjerice na 25 posto veličine okna. Lokalizacijski poligoni $\mathbf{b}_k$ formiraju se na temelju povezanih komponenti grafa $\mathbf{c}_k$ na način da se napravi unija svih okana povezane komponente (redak 8). Valja napomenuti da se povezane komponente s manje od N pridruženih okana uklanjaju iz razmatranja (redak 7). Ocijenjeno je da za takve komponente nema dovoljno dokaza da one ukazuju na prisutnost objekta u slici.

Glavna ideja pristupa zasnovanog na pojedinačnim mjerilima jest spriječiti da se formiraju komponente prostornog grafa povezivanjem pozadinskih okana različitih mjerila. Opisani se pristup primjenjuje za određivanje lokalizacijskih poligona u okviru eksperimentalnog vrednovanja slabo nadzirane lokalizacije prometnih znakova u okviru odjeljka 5.2. Navedeno poglavlje sadrži konkretne vrijednosti i obrazloženja za parametre T, P i N.

Algoritam 1 Određivanje lokalizacijskih poligona na temelju pojedinačnih mjerila

Parametri: T: broj okana za generiranje prostornog grafa, P: uvjet preklapanja, N: minimalni broj okana u povezanoj komponenti grafa

Ulaz: skup $\mathbf{S} = \{x_i, m_i, f(x_i)\}$

i) geometrija slikovnih okana x_i

ii) mjerilo $m_i \in \mathbf{M}$ na kojem je x_i uzorkovan, gdje je **M** skup mjerila

iii) odziv okna $f(x_i)$

Inicijalizacija: sortiraj **S** padajuće u odnosu na $f(x_i)$

1: Iz skupa **S** izdvoji T okana najvećeg odziva u skup **D**
2: **Za** sva mjerila $m \in \mathbf{M}$:
3: Iz **D** izdvoji $\{x_i, m_i, f(x_i) \mid m_i = m\} \rightarrow \mathbf{D}_m$
4: Konstruiraj prostorni graf povezanosti $G = \{\mathbf{c}_k\}$ na temelju $\mathbf{D}_m$:
5: Okna x_i i x_j pripadaju povezanoj komponenti $\mathbf{c}_k$ akko $x_i \cap x_j \geq P$
6: **Za sve** povezane komponente $\mathbf{c}_k \in G$:
7: **Ako** broj okana $x_i \in \mathbf{c}_k \geq N$:
8: Na temelju unije okana $x_i \in \mathbf{c}_k$ odredi opisani pravokutnik $\mathbf{b}_k$

Izlaz: skup poligona lokalizacije $\mathbf{B} = \{\mathbf{b}_k\}$

3.5.2 Određivanje lokalizacijskih poligona na temelju ujedinjene mape odziva preko više mjerila

Pristup temeljen na ujedinjenim mapama odziva preko više mjerila opisan je algoritmom 2. Primjenjuje se u konjunkciji s konvolucijskim značajkama u eksperimentalnom vrednovanju lokalizacije pješačkih prijelaza u okviru odjeljka 5.3.

Glavna motivacija za primjenu ovog pristupa je lokalizacija većih objekata kao što su pješački prijelazi. Eksperimentalnim vrednovanjem ustanovljeno je da okna različitih mjerila mogu doprinijeti stvaranju dovoljno velikih lokalizacijskih poligona. Pristup se zasniva na sjedinjenoj mapi odziva svih piksela u slici $\mathbf{H}(p_j)$. Budući da su odzivi pojedinih okana dobiveni nekom od metoda (3.10), (3.12) ili (3.14) na rezoluciji uzorkovanih lokalnih opisnika, obavlja se naduzorkovanje (engl. *upsampling*) na rezoluciju slike $W_{im} \times H_{im}$ (redak 3). Na taj se način dobiva mapa odziva piksela na pojedinačnom mjerilu $\mathbf{H}(p_j, m)$. Kumulativan odziv određenog piksela p_j dobiva se zbrajanjem odziva preko mapa odziva na pojedinačnim mjerilima (redak 4):

$$\mathbf{H}(p_j) = \sum_{m \in \mathbf{M}} \mathbf{H}(p_j, m) \; . \tag{3.23}$$

Algoritam 2 Određivanje lokalizacijskih poligona na temelju ujedinjenih mapama odziva

Parametri: T: prag vrijednosti odziva piksela

Ulaz: skup $\mathbf{S} = \{x_i, m_i, f(x_i)\}$

i) geometrija slikovnih okana x_i

ii) mjerilo $m_i \in \mathbf{M}$ na kojem je x_i uzorkovan, gdje je $\mathbf{M}$ skup mjerila

iii) odziv okna $f(x_i)$

Inicijalizacija: postavi ujedinjenu mapu odziva piksela u rezoluciji originalne slike $\mathbf{H}(p_j) \in \mathbb{R}^{W_{im} \times H_{im}} = \{0\}$, za svaki piksel p_j

1: **Za sve** mjerila $m \in \mathbf{M}$:

2: Iz $\mathbf{S}$ izdvoji $\{x_i, m_i, f(x_i) \mid m_i = m\}$

3: Obavi naduzorkovanje (engl. *upsampling*) $\{f(x_i)\}$ na rezoluciju $W_{im} \times H_{im} \rightarrow$ mapa odziva na mjerilu m: $\mathbf{H}(p_j, m)$

4: Osvježi $\mathbf{H}(p_j) := \mathbf{H}(p_j) + \mathbf{H}(p_j, m)$

5: Iz $\mathbf{H}(p_j)$ izdvoji $\{p_j \mid \mathbf{H}(p_j) > T\} \rightarrow \mathbf{D}$

6: Na temelju povezanih piksela u $\mathbf{D}$ odredi povezane komponente $\mathbf{C} = \{\mathbf{c}_k\}$

7: **Za sve** $\mathbf{c}_k \in \mathbf{C}$:

8: Na temelju piksela $\{p_j\} \in \mathbf{c}_k$ odredi konveksne ljuske $\mathbf{b}_k$

Izlaz: skup poligona lokalizacije $\mathbf{B} = \{\mathbf{b}_k\}$

Na temelju sjedinjene mape odziva $\mathbf{H}(p_j)$, izdvajaju se pikseli čiji je odziv veći od praga za-

danog parametrom algoritma T (redak 5). U praksi se T postavlja na srednju vrijednost piksela pozitivnog odziva $T = 1/P \cdot \sum_j H(p_j) \mid H(p_j) > 0$, gdje P označava broj takvih piksela. Pikseli koji zadovoljavaju zadani uvjet grupiraju se u povezane komponente, a na temelju piksela pojedinačnih komponenti generiraju se konveksne ljuske (redak 8). Dobivene konveksne ljuske predstavljaju lokalizacijske poligone $\mathbf{b}_k$ i dodaju se u skup $\mathbf{B}$.

3.5.3 Uklanjanje višestrukih poligona lokalizacije

Stvaranjem poligona lokalizacije primjenom obje metode opisane u odjeljku 3.5 mogu nastati višestruke detekcije na području objekta od interesa. Mogućnost nastanka višestrukih lokalizacija vjerojatnija je u slučaju pristupa zasnovanog na pojedinačnom mjerilu gdje postupak može nezavisno generirati međusobno preklapajuće poligone na različitim mjerilima. Budući da se najviše jedan poligon na području objekta računa kao ispravna lokalizacija, a ostali kao lažni pozitivi, potrebno je ugraditi dodatan mehanizam uklanjanja potencijalnih višestrukih detekcija. Postupak uklanjanja višestrukih poligona lokalizacija prikazan je u okviru algoritma 3.

Algoritam 3 Uklanjanje višestrukih poligona lokalizacija

Parametri: M: prag za vrijednost omjera presjeka i unije dva poligona

Ulaz: skup lokalizacijskih poligona $\mathbf{P} = \{p_i, f(p_i), a_i\}$, gdje

i) lokalizacijski poligon p_i

ii) združeni odziv lokalizacijskog poligona $f(p_i)$

iii) površina lokalizacijskog poligona a_i

1: **Funkcija** PRONAĐI ZALIHOSNE POLIGONE(**P**, c)
2: Sortiraj poligone **P** prema kriteriju c
3: Inicijaliziraj listu zalihosnih poligona $\mathbf{Z} = \{\}$
4: **Za sve** poligone $p_i \in \mathbf{P}$:
5: **Za** $p_j \in \mathbf{P} \mid c(p_j) < c(p_i)$:
6: Izračunaj omjer presjeka i unije $IoU(p_i, p_j) = p_i \cap p_j / p_i \cup p_j$
7: **Ako** $IoU(p_i, p_j) > M$:
8: Spremi p_j za uklanjanje: $\mathbf{Z} := [\mathbf{Z}, p_j]$
9: **Funkcija** UKLONI VIŠESTRUKE LOKALIZACIJE(**P**)
10: Korak 1: $\mathbf{Z_1}$ = PRONAĐI ZALIHOSNE POLIGONE(**P**, c = odziv)
11: Ukloni $\{p_j \in \mathbf{Z_1}\}$ iz **P**
12: Korak 2: $\mathbf{Z_2}$ = PRONAĐI ZALIHOSNE POLIGONE(**P**, c = površina)
13: Ukloni $\{p_j \in \mathbf{Z_2}\}$ iz **P**

Izlaz: Pročišćeni skup poligona lokalizacija **P**

Funkcija `ukloni_višestruke_lokalizacije` predstavlja glavnu funkciju algoritma i po-

dijeljena je u dva koraka. U svakom od koraka poziva se funkcija `pronađi_zalihosne_poligone` koja identificira poligone za uklanjanje prema zadanom kriteriju. Inicijalno se primjenjivao samo prvi korak algoritma uklanjanja višestrukih lokalizacija [20]. Rezultati u odjeljku 5.2.2 pokazuju da primjena drugog koraka uz ostale optimizacije povećava lokalizacijsku točnost za 9 posto (86 posto u retku 7 tablice 5.2 naspram 77 posto u retcima 4 i 5 tablice 5.4).

U prvom se koraku (redak 10) poligoni lokalizacija **P** sortiraju padajuće prema združenoj vrijednosti odziva značajki koje su dovele do formiranja poligona, odnosno $c = odziv$ (redak 2). Zatim se slijedno pretražuje lista lokalizacijskih poligona i pronalaze poligoni niže vrijednosti odziva koji se preklapaju s poligonima većeg iznosa odziva (retci 4 i 5). Uvjet preklapanja regulira se pragom $M \in [0 \ldots 1]$ koji označava omjer presjeka i unije (engl. *intersection-over-union criteria*) dvaju poligona. Poligoni nižeg odziva koji zadovoljavaju $IoU(p_i, p_j) > M$ spremaju se u listu kandidata za uklanjanje (retci 7 i 8). U praksi se parametar M najčešće postavlja na vrijednost kriterija valjanosti poligona lokalizacije. U okviru eksperimenata u poglavlju 5 za provjeru valjanosti koristi se omjer presjeka i unije između ručno označenog poligona (engl. *ground truth polygon*) i dobivenog lokalizacijskog poligona [131]. U literaturi [131] se najčešće ta vrijednost postavlja na 0.5, odnosno presjek dvaju poligona čini barem 50 posto njihove unije.

U drugom se koraku (12) preostali lokalizacijski poligoni sortiraju prema padajućoj vrijednosti površine poligona, $c = površina$. Analogno prvom koraku, detektiraju se poligoni manje površine za koje je $IoU(p_i, p_j) > M$ u odnosu na poligone veće površine.

Poglavlje 4

Reprezentacije prostornog rasporeda slikovnih okana

Primjenom rijetkog lokalizacijskog modela, učenog na temelju Fisherovih vektora cjelokupnih slika, broj potencijalnih lokacija traženog objekta znatno se smanjuje. Slika 4.1 ilustrira rezultate primjene takvog modela na primjeru lokalizacije prometnih znakova. Preglednosti radi, slika je prikazana u sivoj boji (engl. *grayscale*), dok su središta slikovnih okana, za koje postupak efikasnog proračuna odziva daje pozitivan ishod, prikazana različitim bojama*. Pri tome se za svako slikovno okno razmatra slikovna riječ (Gaussova komponenta) kojoj odgovara najveća vrijednost vjerojatnosti pridruživanja prema izrazu (2.23), odnosno dominantna slikovna riječ. Boja slikovnog okna označava pripadnost dominatnoj slikovnoj riječi. Slika pokazuje da se primjenom lokalizacijskog pristupa ove disertacije uspješno identificiraju slikovna okna odgovorna za prisutnost traženog objekta u slici. Ipak, u okviru pozadinskih objekata sličnih uzoraka kao u razredu traženog objekta (krovovi kuća) u ovom se slučaju javljaju lažni odzivi. Analiza pokazuje da se dominantne slikovne riječi pojavljuju i na traženom objektu i u pozadini, no njihov međusoban raspored različit je u ta dva slučaja. Jedno od mogućih rješenja problema lažnih pozitiva leži u predstavljanju lokalnog prostornog rasporeda slikovnih riječi. U okviru ovog poglavlja predložena su, stoga, dva tipa reprezentacije lokalnog prostornog rasporeda slikovnih riječi: prostorni histogrami (odjeljak 4.1) i prostorni Fisherovi vektori (odjeljak 4.2).

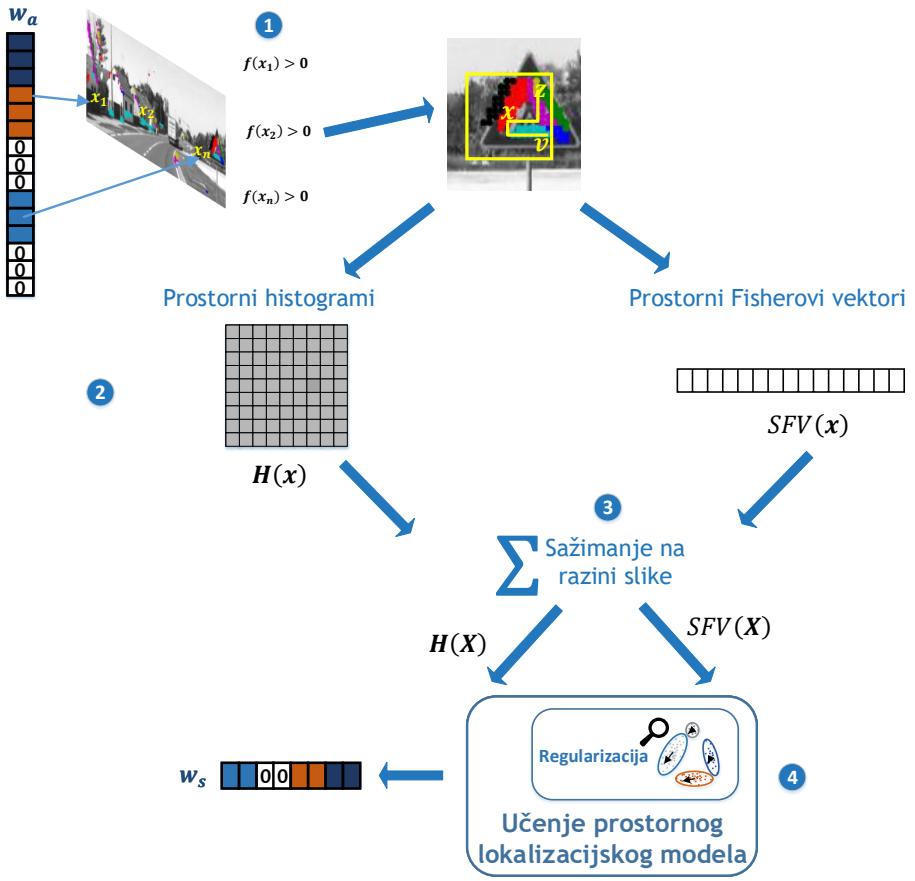
Shematski prikaz procesa učenja prostornog lokalizacijskog modela prikazan je na slici 4.2. Lokalizacija prostornim rasporedom primjenjuje se u drugoj razini sustava za lokalizaciju „odozdo prema gore". Za izgradnju prostornih opisnika, ključna je analiza lokalizacijskog modela opisana u odjeljku 3.4.2. U okviru analize, utvrđuju se diskriminativne slikovne riječi relevantne za modeliranje prostornog rasporeda. Umjesto oznake $\mathbf{S}$ za skup diskriminativnih komponenti modela koja se primjenjuje u odjeljku 3.4.2, ovdje se koristi oznaka $\mathbf{A} = \{a_i\}$, gdje

*Ovu i ostale slike u boji najbolje je pogledati u elektronskoj verziji ove disertacije dostupnoj na adresi http://www.zemris.fer.hr/ ssegvic/zadrija17dr.pdf

Slika 4.1: Motivacija za izgradnju modela prostornog rasporeda: prikazana su središta slikovnih okana za koje postupak efikasnog proračuna odziva (odjeljak 3.4) daje pozitivan ishod. Boja slikovnog okna označava pripadnost dominatnoj slikovnoj riječi (primjerice, slikovna riječ a_1 označena je ljubičastom bojom, a_2 cijan bojom, a a_3 zelenom). Karakteristične slikovne riječi pojavljuju se na traženom objektu (prometni znak) i u pozadini (krovovi, drveće), no prostorni raspored između njih različit je na objektu u odnosu na pozadine. Lokalizacijski poligoni dobiveni na temelju svih pozitivnih okana (odjeljak 3.5.1) označeni su žutim pravokutnicima.

vrijedi $|\mathbf{A}| = K_{\mathrm{w}}$. Cilj promjene notacije je naglasiti razliku između dvije razine lokalizacije „odozdo prema gore". U prvoj se razini (poglavlje 3) za predstavljanje slike i slikovnih okana isključivo koriste značajke kojima se opisuje izgled slike (engl. *appearance*) te se, stoga, lokalizacijski model prve razine označava s $\mathbf{w_a}$, a odgovarajući skup relevantnih slikovnih riječi s $\mathbf{A}$. Na temelju rezultata prve razine lokalizacije, u lokalizaciju prostornim rasporedom propagiraju se slikovna okna pozitivnog odziva (korak (1)). Za svako slikovno okno, razmatra se okruženje dimenzija $W \times H$ za koje se modelira prostorni raspored između okna u središtu i ostalih okana pozitivnog odziva. Prostorni raspored opisuje se primjenom prostornih histograma ili prostornih Fisherovih vektora (korak (2)). Nakon kôdiranja opisnika lokalnih okruženja, obavlja se sažimanje (korak (3)) na razini slike. Na temelju dobivenih opisnika obavlja se učenje rijetkog prostornog modela $\mathbf{w_s}$ (korak 4). Time završava faza učenja prostornog modela lokalizacije. Faza lokalizacije prostornim modelom također podrazumijeva prethodnu primjenu modela izgleda $\mathbf{w_a}$ za utvrđivanje okana pozitivnog odziva. Dobivena slikovna okna se kôdiraju u opisnike prostornog rasporeda i na njih se primjenjuje lokalizacijski model prostornog rasporeda $\mathbf{w_s}$.

Uz spomenuto filtriranje lažnih odziva u okviru teških pozadina, prikazano na slici 4.1, dodatna prednost primjene prostornih modela reprezentacije kao opisnika druge razine jest činjenica da se na taj način smanjuje broj hiperparametara algoritma konstrukcije poligona lokalizacije (odjeljak 3.5). Primjerice, algoritam temeljen na pojedinačnim mjerilima parametriziran je s tri hiperparametra: i) brojem okana T najvišeg odziva koja se koriste kao osnova za izgradnju prostornog grafa, ii) uvjetom preklapanja takvih okana P te naposljetku iii) minimalnim brojem okana N potrebnih za formiranje valjanih poligona lokalizacije. U slučaju prostornog lokaliza-



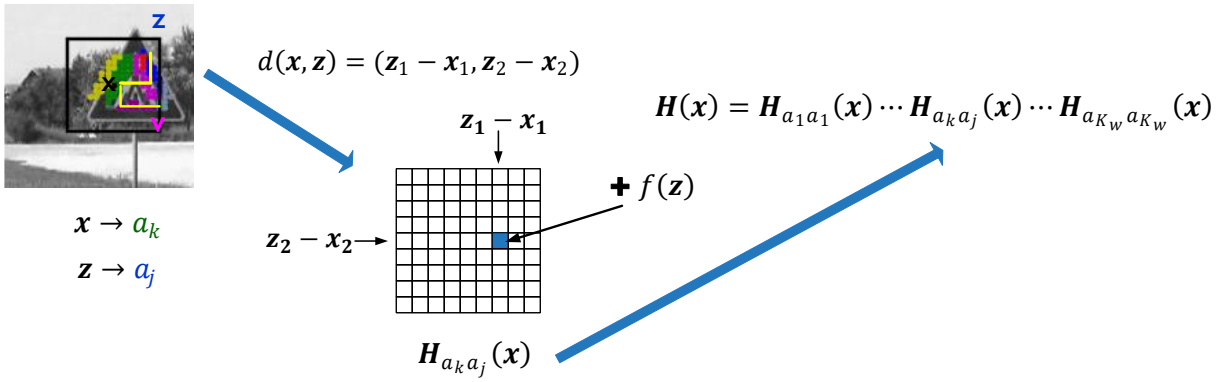
Slika 4.2: Učenje prostornog lokalizacijskog modela.

cijskog modela, lokalizacijski se poligoni konstruiraju na temelju *svih pozitivno klasificiranih okana* od strane modela $\mathbf{w_s}$, čime se uklanja potreba za odabirom hiperparametra T.

Opisani postupak lokalizacije prostornim rasporedom vrednovan je na problemu lokalizacije prometnih znakova opisanom u okviru odjeljka 5.2.3.

4.1 Histogram prostornog rasporeda

Postupak izgradnje prostornih histograma opisan je u okviru algoritma 4, dok slika 4.3 ilustrira korake navedene u okviru algoritma. Za slikovno okno $\mathbf{x}$ pozitivnog odziva $f(\mathbf{x}) > 0$, razmatra se lokalno okruženje dimenzija $W \times H$ (redak 5), gdje su W i H parametri algoritma. Algoritam razmatra sva slikovna okna $\mathbf{z}$ u lokalnom okruženju u odnosu na središnje okno $\mathbf{x}$ (retci 6 i 7). Za par slikovnih okana $(\mathbf{x}, \mathbf{z})$, gdje su odgovarajuće dominantne slikovne riječi pridružene na način $\mathbf{x} \mapsto a_k$ i $\mathbf{z} \mapsto a_j$, najprije se računa vektor relativnog pomaka okna $\mathbf{z}$ u odnosu na središnje okno $\mathbf{x}$ označen s $d(\mathbf{x}, \mathbf{z}) \in \mathbb{R}^2$ (redak 8). Statistika o dobivenom vektoru pomaka pohranjuje se u odgovarajući odjeljak 2D histograma para slikovnih riječi $\mathbf{H}_{a_k a_j}(\mathbf{x})$. Histogrami parova slikovnih riječi diskretizirani su u b odjeljaka (engl. *bin*) u odnosu na obje dimenzije,



Slika 4.3: Ilustracija procesa izgradnje prostornih histograma.

gdje je b hiperparametar algoritma. Doprinos vektora pomaka $d(\mathbf{x},\mathbf{z})$ u odnosu na histogram para riječi $\mathbf{H}_{a_ka_j}(\mathbf{x})$ otežan je s iznosom odziva okna $\mathbf{z}$ označenim s $f(\mathbf{z})$ (redak 9). Konačan opisnik lokalnog okruženja okna $\mathbf{x}$ označen je s $\mathbf{H}(\mathbf{x})$ te se dobiva nadovezivanjem histograma parova riječi $\mathbf{H}_{a_ka_j}(\mathbf{x})$. Dimenzionalnost konačnog opisnika jednaka je $K_\mathrm{w}^2 \cdot b^2$.

Algoritam 4 Izgradnja prostornih histograma

Parametri: $W \times H$: dimenzije lokalnog okruženja, b broj odjeljaka u prostornom histogramu

Ulaz: Skup diskriminativnih slikovnih riječi $\mathbf{A} = \{a_k\}$, $|\mathbf{A}| = K_\mathrm{w}$

Skup slikovnih okana $\mathbf{X} = \{\mathbf{x}, f(\mathbf{x}) > 0, \mathbf{x} \mapsto a_k \mid \max_{a_k} p(a_k|\mathbf{x})\}$, gdje

i) slikovno okno $\mathbf{x}$ sa središtem u $(\mathbf{x}_1, \mathbf{x}_2)$

ii) odziv okna u odnosu na lokalizacijski model prve razine $\mathbf{w_a}$: $f(\mathbf{x})$

iii) slikovna riječ a_k kojoj je $\mathbf{x}$ dodijeljen

1: **Za** sva okna $\mathbf{x} \in \mathbf{X}$:

2: Inicijaliziraj prostorni histogram okna $\mathbf{H}(\mathbf{x}) \in \mathbb{R}^{K_\mathrm{w}^2 \cdot b^2} = [\,]$

3: **Za** sve parove riječi (a_k, a_j) :

4: Incijaliziraj histograme parova riječi $\mathbf{H}_{a_ka_j}(\mathbf{x}) \in \mathbb{R}^{b^2} = \{0\}$

5: Odredi lokalno okruženje oko $(\mathbf{x}_1, \mathbf{x}_2)$ dimenzija $W \times H$

6: Pronađi sva okna $\mathbf{Z} = \{\mathbf{z}, f(\mathbf{z}) > 0, \mathbf{z} \mapsto a_j \mid \max_{a_j} p(a_j|\mathbf{z})\}$ u lokalnom okruženju

7: **Za** okna u lokalnom okruženju $\mathbf{z} \in \mathbf{Z}$:

8: Izračunaj vektore pomaka $d(\mathbf{x},\mathbf{z}) = (\mathbf{z}_1 - \mathbf{x}_1,\ \mathbf{z}_2 - \mathbf{x}_2)$

9: Osvježi histogram $\mathbf{H}_{a_ka_j}(\mathbf{x})[d(\mathbf{x},\mathbf{z})] = \mathbf{H}_{a_ka_j}(\mathbf{x})[d(\mathbf{x},\mathbf{z})] + f(\mathbf{z})$

10: **Za** sve parove riječi (a_k, a_j) :

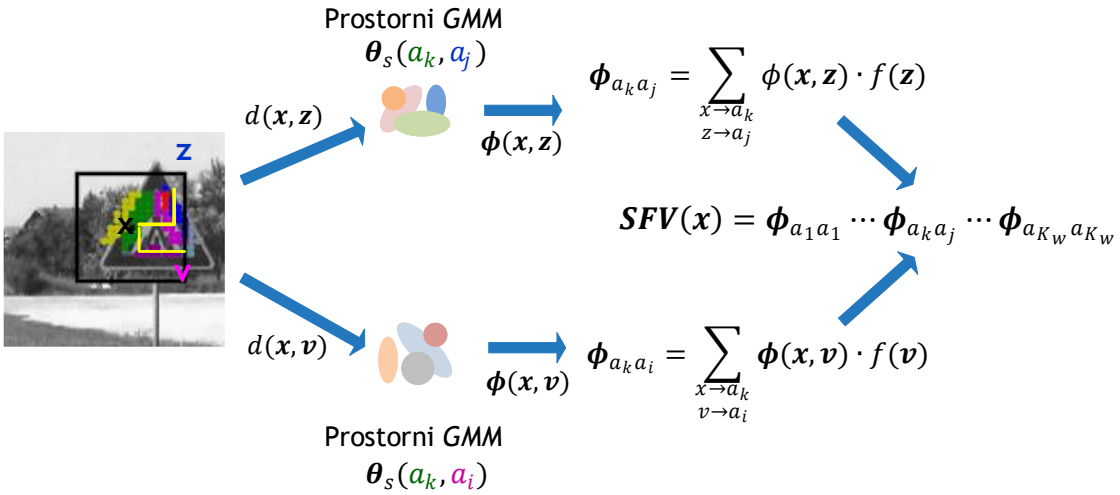
11: Osvježi $\mathbf{H}(\mathbf{x}) = [\mathbf{H}(\mathbf{x}),\ \mathbf{H}_{a_ka_j}(\mathbf{x})]$

12: Dodaj $\mathbf{H}(\mathbf{x})$ u skup prostornih histograma $\mathbf{H}$

Izlaz: Skup prostornih histograma $\mathbf{H}$

4.2 Prostorni Fisherov vektor

Izgradnja prostornih Fisherovih vektora opisana je u okviru algoritma 5 i slike 4.4. Slično kao i u slučaju prostornih histograma, proces izgradnje započinje izvođenjem vektora relativnih pomaka slikovnog okna u središtu okruženja u odnosu na ostala okna. Ključna razlika jest da se prostorni odnos između parova slikovnih riječi (a_k, a_j) ne modelira prostornim histogramom, već Fisherovim vektorom u odnosu na prostorni model raspodjele Gaussovih mješavina. Prostorni model mješavine Gaussovih raspodjela parametriziran je sa $\boldsymbol{\theta}_s(a_k, a_j) = \{\alpha_s, \boldsymbol{\mu}_s, \boldsymbol{\sigma}_s\}_{i=1}^{K_\mathrm{s}},\ \alpha_s \in \mathbb{R},\ \boldsymbol{\mu}_s \in \mathbb{R}^2,\ \boldsymbol{\sigma}_s \in \mathbb{R}^2$. Parametar K_s označava broj komponenti prostornog GMM-a. Slično kao i kod modela mješavine Gaussovih raspodjela kojim u prvoj razini lokalizacije modeliramo opisnike slikovnih okana, ovdje se također podrazumijevaju dijagonalne matrice kovarijacije $\boldsymbol{\sigma}_s$. Sasvim općenito, prostorni se GMM može naučiti na postupcima nenadziranog učenja na slučajno odabranom skupu uzoraka vektora pomaka za svaki par slikovnih riječi (a_k, a_j). U eksperimentima predstavljenim u okviru odjeljka 5.2.3 koristi se prostorni GMM dijeljen između svih parova slikovnih riječi $\boldsymbol{\theta}_s(a_1, a_2) = \ldots = \boldsymbol{\theta}_s(a_k, a_j) \ldots \boldsymbol{\theta}_s(a_{K_\mathrm{w}}, a_{K_\mathrm{w}})$. Srednja vrijednost $\boldsymbol{\mu}_s$ i varijanca $\boldsymbol{\sigma}_s$ pojedinih komponenti prostornog GMM-a odgovaraju srednjoj vrijednosti i varijanci uniformne raspodjele u četiri kvadranta kvadrata jedinične površine [108]. Vektori pomaka $d(\mathbf{x}, \mathbf{z})$ kôdiraju se u odnosu na prostorni GMM, a kao rezultat dobivaju se Fisherovi vektori $\boldsymbol{\phi}(\mathbf{x}, \mathbf{z}) \in \mathbb{R}^{K_\mathrm{s}(2\cdot 2+1)}$ (redak 9). Dobiveni Fisherovi vektori agregiraju se u odgovarajući Fisherov vektor koji predstavlja par slikovnih riječi $\boldsymbol{\phi}_{a_k a_j}$. Slično kao i kod prostornih histograma, doprinos pojedinog Fisherova vektora otežan je odzivom okna $\mathbf{z}$ označenim s $f(\mathbf{z})$ (redak 10). Konačan opisnik lokalnog okruženja dobiva se nadovezivanjem opisnika svih parova slikovnih riječi $\boldsymbol{\phi}_{a_k a_j}$ (retci 11 i 12). Dimenzionalnost prostornog Fisherova vektora odgovara $K_\mathrm{w}^2 \cdot K_\mathrm{s} \cdot (2 \cdot 2 + 1) = K_\mathrm{w}^2 \cdot K_\mathrm{s} \cdot 5$.



Slika 4.4: Ilustracija procesa izgradnje prostornih Fisherovih vektora.

Algoritam 5 Izgradnja prostornih Fisherovih vektora

Parametri: $W \times H$: dimenzije lokalnog okruženja, K_s broj komponenata prostornog rječnika

Ulaz: Skup diskriminativnih slikovnih riječi $\mathbf{A} = \{a_k\}$, $|\mathbf{A}| = K_\mathrm{w}$

Skup slikovnih okana $\mathbf{X} = \{\mathbf{x}, f(\mathbf{x}) > 0, \mathbf{x} \mapsto a_k \mid \max_{a_k} p(a_k|\mathbf{x})\}$, gdje

i) slikovno okno $\mathbf{x}$ sa središtem u $(\mathbf{x}_1, \mathbf{x}_2)$

ii) odziv okna u odnosu na lokalizacijski model izgleda $\mathbf{w_a}$: $f(\mathbf{x})$

iii) slikovna riječ a_k kojoj je $\mathbf{x}$ dodijeljen

Za sva okna $\mathbf{x} \in \mathbf{X}$:

  Inicijaliziraj prostorni Fisherov vektor okna $\mathbf{SFV}(\mathbf{x}) \in \mathbb{R}^{K_\mathrm{w}^2 \cdot K_\mathrm{s} \cdot 5} = [\,]$

  Za sve parove riječi (a_k, a_j) :

    Incijaliziraj Fisherove vektore parova riječi $\boldsymbol{\phi}_{a_k a_j} \in \mathbb{R}^{K_\mathrm{s} \cdot 5} = \{0\}$

  Odredi lokalno okruženje oko $(\mathbf{x}_1, \mathbf{x}_2)$ dimenzija $W \times H$

  Pronađi sva okna $\mathbf{Z} = \{\mathbf{z}, f(\mathbf{z}) > 0, \mathbf{z} \mapsto a_j \mid \max_{a_j} p(a_j|\mathbf{z})\}$ u lokalnom okruženju

  Za okna u lokalnom okruženju $\mathbf{z} \in \mathbf{Z}$:

    Izračunaj vektore pomaka $d(\mathbf{x}, \mathbf{z}) = (\mathbf{z}_1 - \mathbf{x}_1,\ \mathbf{z}_2 - \mathbf{x}_2)$

    Obavi kôdiranje $d(\mathbf{x}, \mathbf{z})$ u odnosu na prostorni GMM $\boldsymbol{\theta}_s(a_k, a_j) \to \boldsymbol{\phi}(\mathbf{x}, \mathbf{z})$

    Osvježi $\boldsymbol{\phi}_{a_k a_j} = \boldsymbol{\phi}_{a_k a_j} + f(\mathbf{z}) \cdot \boldsymbol{\phi}(\mathbf{x}, \mathbf{z})$

  Za sve parove riječi (a_k, a_j) :

    Osvježi $\mathbf{SFV}(\mathbf{x}) = [\mathbf{SFV}(\mathbf{x}),\ \boldsymbol{\phi}_{a_k a_j}]$

  Dodaj $\mathbf{SFV}(\mathbf{x})$ u skup prostornih Fisherovih vektora $\mathbf{SFV}$

Izlaz: Skup prostornih histograma $\mathbf{SFV}$

4.2.1 Optimizacije izračuna prostornih Fisherovih vektora

Za zadanu dimenzionalnost lokalnog okruženja $W \times H$ i okno u središtu, može se izdvojiti konačan broj vektora relativnih pomaka $d(\mathbf{x}, \mathbf{z})$ koji veličinom odgovara površini okruženja. Uz poznate prostorne modele Gaussovih mješavina za parove riječi $\boldsymbol{\theta}_s(a_k, a_j)$, Fisherovi vektori $\boldsymbol{\phi}(\mathbf{x}, \mathbf{z})$ se mogu unaprijed izračunati i spremiti u priručnu tablicu (engl. *lookup table*). Ukupan broj Fisherovih vektora u priručnoj tablici odgovara $K_{\mathrm{w}}^2 \cdot W \cdot H$. Prednost navedene optimizacije leži u činjenici da se skupe operacije proračuna gradijenata u odnosu na parametre prostornog Gaussova modela ne moraju računati u fazi lokalizacije, već se dobiveni Fisherovi vektori samo otežavaju odgovarajućim odzivima i agregiraju u opisnike parova.

4.3 Rasprava prostornih opisnika

U okviru ovog poglavlja predstavljena su dva tipa opisnika za modeliranje lokalnih prostornih odnosa između parova slikovnih riječi. U nastavku su sažete glavne značajke oba opisnika te njihove prednosti i nedostaci.

Prostorni histogrami su jednostavniji u odnosu na prostorne Fisherove vektore jer proces kôdiranja uključuje jednostavnu kvantizaciju vektora pomaka u odgovarajuće odjeljke histograma. S druge strane, kod prostornih Fisherovih vektora, vektori pomaka projiciraju se u prostor Fisherovih vektora primjenom nelinearne Fisherove jezgre u odnosu na prostorni GMM. Navedeni nedostatak prostornih Fisherovih vektora može se zaobići primjenom optimizacija predstavljenih u odjeljku 4.2.1, koje uključuju proračun Fisherovih vektora i spremanje u odgovarajuću preglednu tablicu prije samog procesa testiranja.

S obzirom na kriterij ekspresivne moći prostornih opisnika, Fisherovi vektori omogućavaju vjernije predstavljanje prostornog rasporeda. Nedostatak prostornih histograma leži u nemogućnosti da se modelira raspodjela vektora pomaka unutar pojedinog odjeljka histograma. S druge strane, prostorni Fisherovi vektori temelje se se na generativnom modelu opisanom funkcijom gustoće vjerojatnosti. Kôdiranjem u odnosu na GMM bilježi se kako pojedini vektor pomaka utječe na promjenu parametara generativnog modela. Na taj se način vjernije predstavljaju razlike u lokalnom prostornom rasporedu.

Poglavlje 5

Eksperimentalni rezultati

Izvedba eksperimentalnog sustava obuhvaća vrednovanje sustava za lokalizaciju „odozdo prema gore" opisanog u okviru poglavlja 3 i modela prostornog rasporeda slikovnih riječi opisanog u poglavlju 4. Iscrpno vrednovanje funkcionalnosti provedeno je na konkretnim problemima iz stvarnog života:

- automatizaciji pregleda cestovne infrastrukture (odjeljak 5.2)
- automatizaciji digitalne kartografije primjenom masovno prikupljenih podataka (odjeljak 5.3.

Automatizacija pregleda cestovne infrastrukture ima jasan potencijal za povećanje sigurnosti cestovnog prometa [53]. U ovom radu razmatraju se trokutni znakovi opasnosti kao prioritetni element prometne signalizacije. Unatoč standardiziranom izgledu, prometne znakove nije lako lokalizirati, posebice modelima dobivenim slabo nadziranim učenjem. Glavni izazov predstavlja njihova veličina koja u okviru korištenog skupa podataka [72] varira od dva promila do šest posto površine slike (u prosjeku oko jedan posto). U okviru ovog poglavlja, najprije se razmatra problem lokalizacije prometnih znakova u odnosu na: i) tipove normalizacija Fisherovih vektora opisane u odjeljku 3.2 kao i ii) različite tipove regularizacija opisane u odjeljku 3.3. Provedena je iscrpna vremenska analiza izvođenja pojedinih koraka predstavljenoga pristupa. Rezultati opisani u odjeljku 5.2.3 [20] opisuju vrednovanje prostornih modela reprezentacije (poglavlje 4). Opisani rezultati datiraju kronološki ranije u odnosu na one predstavljene u odjeljku 5.2.2 te ih odlikuje slabija lokalizacijska točnost. Razlog tome su: i) uži interval pretraživanja regularizacijskog parametra λ u okviru unakrsne provjere (engl. *cross validation*) te ii) za uklanjanje višestrukih poligona lokalizacija primijenjen je isključivo prvi korak algoritma 3 opisan u odjeljku 3.5.3. Motivacija i potencijalni izazovi procesa utvrđivanja ispravnosti cestovne infrastrukture prethodno su opisani u okviru odjeljka 1.2.1.

Zadatak automatizacije digitalne kartografije razmatra se u odjeljku 5.3 na primjeru slabo nadzirane lokalizacije pješačkih prijelaza. Kao i u slučaju prometnih znakova, navode se eksperimenti iz područja klasifikacije slika. Detaljno se razmatraju različiti tipovi normalizacija

Fisherova vektora te tipova regularizacija uz odgovarajuću vremensku analizu. Za generiranje oznaka na razini slike koriste se dobrovoljno prikupljeni podaci iz OpenStreetMap karte [62]. Prednosti automatizacije procesa kartiranja navedene su u odjeljku 1.2.2.

5.1 Mjere vrednovanja

U okviru eksperimenata predstavljenih u ovom poglavlju razmatraju se problemi binarne klasifikacije (i lokalizacije) primjenom linearnih diskriminativnih modela. Za potrebe definiranja odgovarajućih mjera vrednovanja, klasifikacijski ishod * (engl. *classification score*) pojedinog primjera (slike ili slikovnog okna) označava se sa $f(\mathbf{x})$, a pripadajući prag klasifikatora sa b. Primjer se klasificira pozitivno ako vrijedi $f(\mathbf{x}) > b$.

Kao mjera klasifikacijske i lokalizacijske točnosti navodi se prosječna preciznost (engl. *average precision*) *AP* [34, 131]. Prosječna preciznost označava površinu ispod krivulje koja prikazuje odnos između odziva (engl. *recall*) na apscisi i preciznosti (engl. *precision*) na ordinati (*PR* krivulje) [34, 153]. Slika 5.1 ilustrira primjer *PR* krivulje, gdje je dobivena prosječna preciznost *AP* u iznosu od 92 posto. Krivulja preciznosti konstruira se na način da se dobiveni ishodi klasifikacije $f(\mathbf{x})$ sortiraju prema veličini: od većih ka manjima. Shodno sortiranim vrijednostima ishoda klasifikacije, varira se prag klasifikatora b. Svaka točka krivulje odgovara određenoj vrijednosti praga b, pri čemu krajnje lijeva točka odgovara najvećem iznosu praga, a krajnje desna najmanjem. Odziv R (vrijednost na apscisi) u određenoj točki krivulje definira se kao udio ispravno klasificiranih pozitivnih primjera u odnosu na ukupan broj pozitivnih primjera s obzirom na vrijednost praga b [153]:

$$R = \frac{T_p}{T_p + F_n} \,. \tag{5.1}$$

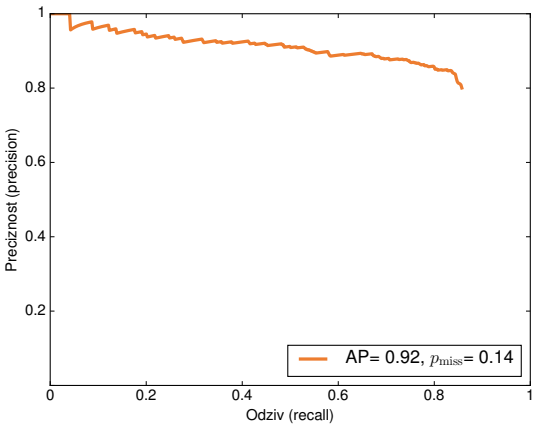
Parametar T_p označava broj ispravno klasificiranih pozitivnih primjera, a F_n broj promašaja. Preciznost P (vrijednost na ordinati) definira se kao udio ispravno klasificiranih pozitivnih primjera u odnosu na sve primjere pozitivnog ishoda klasifikacije [153]:

$$P = \frac{T_p}{T_p + F_p} \,. \tag{5.2}$$

Parametar F_p označava broj lažnih pozitiva, odnosno negativnih primjera za koje je dobiveni rezultat klasifikacije veći od praga za tu točku krivulje.

Kako bi se lokalizacijska performansa okarakterizirala *AP*-om, potrebno je definirati uspješnost pojedinog lokalizacijskog ishoda (engl. *loacalization score*). Taj kriterij treba odražavati

*Kako bi se izbjegla potencijalna konfuzija u terminologiji, u okviru ovog odjeljka se umjesto termina „klasifikacijski odziv" (engl. *classification score*) koristi termin „klasifikacijski ishod". Termin „odziv" koristi se za označavanje odziva (engl. *recall*) kao mjere vrednovanja.



Slika 5.1: Primjer krivulje preciznosti u odnosu na odziv (engl. *precision-recall curve*). Prikazana je prosječna preciznost *AP* u iznosu od 92 posto. Krivulja je konstruirana na temelju zbirnih odziva generiranih lokalizacijskih poligona. Za 14 posto objekata nisu formirani lokalizacijski poligoni. Shodno tome, nije postignut maksimalni odziv $R = 1$, te je zabilježena frekvencija promašaja p_{miss} u iznosu od 0.14.

poklapanje opisanog poligona kojeg je proizveo lokalizacijski postupak B_p s točnim poligonom koji je ručno označen za potrebe vrednovanja B_{gt}. Obično se u tu svrhu koristi omjer presjeka i unije tih dvaju poligona (engl. *Intersection over Union, IoU*) [131]:

$$\frac{area(B_p \cap B_{gt})}{area(B_p \cup B_{gt})} > IoU \ . \tag{5.3}$$

Parametar *IoU* označava granično preklapanje za koje se lokalizacija smatra uspješnom. U okviru eksperimentalnog vrednovanja lokalizacije prometnih znakova koristi se granična vrijednost $IoU = 0.5$ prema [131]. Za potrebe vrednovanja lokalizacije pješačkih prijelaza napravljena je detaljna analiza performansi modela za $IoU \in \{0.1, 0.2, 0.3, 0.4, 0.5\}$.

Za mjerenje lokalizacijske točnosti, uz prosječnu preciznost *AP* također se vrednuje i frekvencija promašaja p_{miss} u krajnje desnoj točki *PR* krivulje:

$$p_{\text{miss}} = \frac{F_n}{T_p + F_n} \ . \tag{5.4}$$

Ovdje izraz F_n označava broj objekata za koje nije generiran poligon lokalizacije, a izraz $T_p + F_n$ ukupan broj ručno označenih objekata. Uspješnost zadataka klasifikacije i lokalizacije dodatno se karakterizira: i) mjerama rijetkosti klasifikacijskog modela te ii) trajanjem pojedinih koraka obrade. U okviru eksperimentalnog vrednovanja, za mjerenje rijetkosti modela koriste se sljedeće veličine:

- K_{w} označava broj komponenti modela $\mathbf{w}$ čija je norma različita od ništice, gdje pojedina komponenta odgovara slikovnoj riječi.
- *ACD* označava prosječnu gustoću komponenti modela, odnosno udio odabranih kompo-

nenti K_{w} u odnosu na ukupan broj komponenti slikovnog rječnika K:

$$ACD = \frac{K_{\mathrm{w}}}{K} \,. \tag{5.5}$$

- *AOD* označava prosječnu gustoću cjelokupnog modela, odnosno udio koeficijenata modela w_i različitih od ništice u odnosu na ukupnu dimenzionalnost modela $|\mathbf{w}|$:

$$ACD = \frac{\sum_i \mathbf{1}\{w_i \neq 0\}}{|\mathbf{w}|} \,. \tag{5.6}$$

Izraz $\mathbf{1}\{w_i \neq 0\} \rightarrow \{0,1\}$ označava indikatorsku funkciju čija je vrijednost jednaka jedinici ukoliko je uvjet $w_i \neq 0$ zadovoljen, dok izraz $|\mathbf{w}|$ označava ukupan broj koeficijenata modela $\mathbf{w}$. Mjera K_{w} pogodna je za slučajeve kada se žele pokazati fine razlike u rijetkosti različitih modela. S druge strane, mjera *ACD* se koristi kada želimo pokazati razlike u rijetkostima modela učenih nad Fisherovim vektorima različitih veličina GMM rječnika.

Brzina izvođenja razvijenih postupaka karakterizirana je sljedećim vremenima:

- t_{lf} označava vrijeme potrebno za izvođenje opisnika niske razine.
- t_{sa} označava vrijeme potrebno za proračun odgovornosti da neka Gaussova komponenta generira pojedini lokalni opisnik prema jednadžbi (2.23).
- t_{op} označava vrijeme potrebno za proračun odziva Fisherovih vektora okana dodijeljenih odabranim komponentama modela prema jednadžbama (3.10), (3.12) ili (3.14).

5.2 Lokalizacija prometnih znakova

U okviru eksperimenata predstavljenih u ovom odjeljku naglasak je na lokalizaciji trokutnih znakova opasnosti. Prema bečkoj konvenciji [154] definirana su 33 različita tipa znakova opasnosti. Za potrebe učenja lokalizacijskog modela i testiranja, koristi se skup podataka TS2010a[†] [72]. Navedeni skup podataka sadrži 3296 slika izvedenih iz video zapisa prikupljenog na području lokalnih cesta sjeverne Hrvatske. Video zapis snimljen je kamerom montiranom na vozilo, a rezolucija dobivenih slika iznosi 720×576. Slike su podijeljene u dva međusobno disjunktna podskupa: i) podskup za učenje od 1705 slika te ii) podskup za testiranje od 1591 slika. Sve slike određenog fizičkog znaka nalaze se u istom podskupu. Podskup za učenje sadrži 453 slike s trokutnim prometnim znakovima (pozitivni primjeri za učenje), dok podskup za testiranje sadrži 379 slika s jednim prometnim znakom te 60 slika s dva prometna znaka.

Skup podataka TS2010a razlikuje se u odnosu na popularne skupove za lokalizaciju objekata [131] na uslijed sljedećih značajki: i) objekti traženog razreda u prosjeku čine manje od 1 posto ukupne površine slike, ii) pozitivne i negativne slike imaju gotovo jednake pozadine. Zbog

[†]Skup podataka TS2010a može se preuzeti sa `http://multiclod.zemris.fer.hr/ts2010a.shtml`.

navedenih obilježja ocijenjeno je da je lokalizacija prometnih znakova zahtjevan i zanimljiv problem unatoč relativnoj jednostavnosti objekta od interesa.

5.2.1 Detalji izvedbe

Za potrebe reprezentacije slika primjenjuje se gusto uzorkovanje kvadratnih slikovnih okana na različitim mjerilima opisano u odjeljku 2.1.1. Dimenzije izdvojenih slikovnih okana odgovaraju 16, 24, 32 i 40 piksela po dužini i širini, a uzorkovana su s pomakom od 1/8 dužine okna na odgovarajućem mjerilu. Za opisivanje slikovnih okana koriste se SIFT opisnici fiksne orijentacije na svakom od četiri mjerila (engl. *dense SIFT*) [88], čiji se izračun obavlja se pomoću *VLFeat* biblioteke [102]. SIFT opisnici metrički se normaliziraju i nad njima se obavlja dekorelacija algoritmom analize glavnih komponenti (engl. *Principal component analysis, PCA*). Rezultat dekorelacije su opisnici dimenzije $D = 80$.

Na slučajno odabranom uzorku od 10^6 SIFT opisnika izdvojenih iz slika iz podskupa za treniranje, primijenjen je algoritam maksimizacije očekivanja (engl. *expectation-maximization algorithm, EM*) kako bi se naučio model mješavine Gaussovih razdiobi (GMM). Za učenje se koristi implementacija EM algoritma iz programske biblioteke *Yael* [155]. Prilikom učenja podrazumijevaju se sljedeći parametri: i) broj komponenti GMM modela postavljen je na $K = 1024$ te ii) Gaussove komponente karakterizirane su dijagonalnim matricama kovarijacije.

Kôdiranje lokalnih opisnika u Fisherove vektore obavlja se prema jednadžbama (2.24), (2.25), (2.26). Dimenzionalnost rezultatnih Fisherovih vektora iznosi $K \cdot (2 \cdot D + 1) = 1024 \cdot (2 \cdot 80 + 1) = 164864$.

Lokalizacijski se model $\mathbf{w}$ uči na temelju Fisherovih vektora slika $\mathbf{X} = \{\mathbf{X}_i\}$, $\mathbf{X}_i \in \mathbb{R}^{K(2D+1)}$ i odgovarajućih oznaka prisutnosti objekata u slici $\mathbf{y} = \{y_i\}$, gdje $y_i \in \{-1, 1\}$. Prilikom učenja, minimizira se regularizirani logistički gubitak koji odgovara negativnoj log-izglednosti logističke funkcije:

$$\ell(\mathbf{w}, \mathbf{X}, \mathbf{y}) = \frac{1}{N} \sum_{i=1}^{N} \log\left(1 + \exp(-y_i \cdot \mathbf{w}^\top \mathbf{X}_i)\right) + \lambda \cdot \mathscr{R}(\mathbf{w}) \tag{5.7}$$

Parametar N označava broj slika u podskupu za učenje. Kako bi se smanjio utjecaj prenaučenosti, evaluiraju se različiti tipovi funkcija regularizacije $\mathscr{R}(\mathbf{w})$ opisani u odjeljku 3.3.1, pri čemu se vrijednost regularizacijskog parametra određuje putem unakrsne provjere sa 10 odjeljaka (engl. *10-fold cross-validation*). Za učenje modela koristi se programska biblioteka *SPAMS* [156].

Za određivanje poligona lokalizacije koristi se pristup temeljen na pojedinačnom mjerilu opisan u okviru odjeljka 3.5.1. Navedeni pristup je odabran iz razloga što su preliminarni eksperimenti sa slikovnim oknima različitih mjerila rezultirali s lošijom lokalizacijskom točnošću.

U svakoj slici se odabire $T = 100$ slikovnih okana najvećeg odziva te se na temelju njih konstruira graf prostorne povezanosti. Smatra se da su dva slikovna okna povezana ako se preklapaju više od $P \geq 25$ posto u odnosu na površinu slikovnog okna na odgovarajućem mjerilu. Povezane komponente grafa s manje $N < 10$ dodijeljenih slikovnih okana uklanjaju se iz razmatranja.

5.2.2 Vrednovanje modela temeljenih na izgledu

U okviru ovog odjeljka razmatra se utjecaj normalizacija Fisherova vektora te različitih tipova regularizacijskih funkcija za zadatke klasifikacije slike i lokalizacije trokutastih prometnih znakova.

Rezultati klasifikacije slika

Uspješnost modela za klasifikaciju slika prikazana je u tablici 5.1. Eksperimenti su podijeljeni u tri skupine, a u okviru svake od njih vrednuju se redom ℓ_2-gusta regularizacija definirana preko (3.6), rijetka ℓ_1 regularizacija definirana preko (3.7), te rijetka $\ell_{2,1}$ regularizacija definirana jednadžbom (3.8):

- Retci 1 – 3 prikazuju rezultate za modele učene na nenormaliziranim Fisherovim vektorima slika gdje je klasifikacijski odziv dobiven skalarnim produktom Fisherova vektora slike i modela.
- Retci 4 – 6 daju pregled rezultata za modele učene nad Fisherovim vektorima normaliziranim potenciranjem (oznaka p) (3.1) te nakon toga globalnom metričkom normalizacijom (oznaka ℓ_2 global) (3.2), odzivi klasifikacije dobivaju se skalarnim produktom modela i normaliziranog FV slike.
- Retci 7 – 9 prikazuju vrednovanje modela učenih nad Fisherovim vektorima normaliziranim potenciranjem (oznaka p) (3.1) te nakon toga metričkom normalizacijom na razini komponente (oznaka ℓ_2 intra) (3.3), odzivi klasifikacije dobivaju se skalarnim produktom modela i normaliziranog FV slike.

Kao mjere vrednovanja koriste se prosječna preciznost na skupu za učenje (*AP* učenja) i testiranje (*AP* testiranje) te mjere cjelokupne gustoće modela *AOD* i gustoće s obzirom na broj komponenti modela *ACD*. Rezultati pokazuju da rijetki modeli (ℓ_1 i $\ell_{2,1}$ regularizacije) postižu bolje rezultate u odnosu na gusti ℓ_2 model za sve grupe eksperimenata. Najveća razlika u uspješnosti zapažena je u prvoj grupi eksperimenata, gdje po komponentama rijedak $\ell_{2,1}$ model postiže za 9 posto bolji *AP* na skupu za testiranje (redak br. 3 naspram retka br. 1). Primjenom normalizacije potenciranjem i globalne metričke normalizacije, smanjuje se razlika u učinkovitosti rijetkih i gustih modela na 6 posto, a primjenom normalizacije na razini komponente na 3 posto. Nelinearne normalizacije pospješuju učinkovitost klasifikacije za sve tipove modela te se uz normalizaciju potenciranjem, globalnu metričku normalizaciju i $\ell_{2,1}$ regularizaciju

Tablica 5.1: Vrednovanje binarne klasifikacije slika prometnih znakova s obzirom na različite normalizacije Fisherova vektora slike (*p*: potenciranje, ℓ_2 global: metrička na razini cjelokupnog FV, ℓ_2 intra: metrička na razini komponente) i tipove regularizacija (ℓ_2, ℓ_1 i $\ell_{2,1}$). Mjera *AOD* (engl. *average overall density*) označava postotak koeficijenata modela različitih od ništice (od ukupno 164865). Mjera *ACD* (engl. *average component density*) označava udio komponenti modela s normom različitom od ništice K_w/K, od ukupno $K = 1024$. Sve prikazane mjere učinkovitosti izražene su u postocima.

Br.	FV normalizacija	Regularizacija	*AOD*	*ACD*	*AP* učenja	*AP* testiranja
1	-	ℓ_2	92.8	100	100	66
2	-	ℓ_1	0.1	6.1	87	71
3	-	$\ell_{2,1}$	1.0	1.1	83	75
4	p, ℓ_2 global	ℓ_2	92.8	100	100	75
5	p, ℓ_2 global	ℓ_1	0.1	3.8	83	78
6	p, ℓ_2 global	$\ell_{2,1}$	1.1	1.1	87	**81**
7	p, ℓ_2 intra	ℓ_2	92.8	100	100	77
8	p, ℓ_2 intra	ℓ_1	0.1	5.1	84	78
9	p, ℓ_2 intra	$\ell_{2,1}$	0.8	0.8	85	80

dobiva najbolji rezultat klasifikacije (81 posto *AP* na skupu za testiranje). Konfiguracija s metričkom normalizacijom po komponentama (redak 9) daje marginalno lošiji rezultat (80 posto *AP* testiranja), no dobiveni model je nešto rjeđi (za oko 0.3 posto u odnosu na redak 6).

Usporedba rijetkih modela (ℓ_1 naspram $\ell_{2,1}$) pokazuje da rijetki $\ell_{2,1}$ modeli postižu bolju klasifikacijsku točnost na skupu za testiranje (do 4 posto *AP* testiranja), a istovremeno ih karakterizira do 5 puta manja gustoća komponenti (*ACD* mjera) u odnosu na ℓ_1 modele. Bolja klasifikacijska točnost uz veći stupanj rijetkosti modela posljedica je činjenice da $\ell_{2,1}$ regularizacija uzima u obzir specifičnu blokovsku strukturu Fisherovih vektora te efikasno obavlja probir diskriminativnih slikovnih riječi. Sasvim općenito, ℓ_1 modeli postižu veći stupanj rijetkosti na razini cjelokupnog modela, gdje u prosjeku *AOD* = 0.1% koeficijenata modela ima vrijednost različitu od ništice, dok kod $\ell_{2,1}$ modela ta mjera iznosi oko 1%. Međutim, mjera gustoće komponenti pokazuje da su koeficijenti ℓ_1 modela u pravilu raspoređeni preko većeg broja komponenti (*ACD* u rasponu od 3.8 - 6.1%), u odnosu na $\ell_{2,1}$ modele gdje se koristi oko 1% ukupnog broja komponenti slikovnog rječnika. Drugim riječima, $\ell_{2,1}$ modeli su efikasniji za izvođenje jer koriste manji broj komponenti modela.

Zaključno, prilikom klasifikacije prometnih znakova postignuta je maksimalna preciznost od 81 posto (redak 6), što znači da je 19 posto slika neispravno klasificirano. Potencijalno objašnjenje dobivenog rezultata leži u činjenici da pozitivne i negativne slike karakteriziraju gotovo jednake pozadine. Drugim riječima, kontekst ne unosi dodatnu informaciju koja bi

pripomogla zadatku klasifikacije.

Rezultati lokalizacije objekata

Kvalitativni rezultati lokalizacije prometnih znakova prikazani su u okviru tablice 5.2.

Budući da su za skup podataka TS2010a dostupne oznake lokacija objekata na podskupu za treniranje, kao referentni eksperiment prikazan je strogo nadzirani pristup temeljen na histogramima orijentacije gradijenata (engl. *Histogram of Oriented Gradients, HOG*)[4] (redak 1). Kako bi se postigla što bolja lokalizacijska preciznost, HOG značajke izdvojene su primjenom Python sučelja *OpenCV* biblioteke [157] na ukupno 64 mjerila u rasponu od 24×24 do 160×160 (shodno veličini objekata u skupu podataka), uz faktor skaliranja od 1.03 i pomak od 2 piksela. Iz navedenog razloga nije bilo moguće izmjeriti vremena t_{op} i t_{lf} odvojeno te je u tablici dana ukupna vrijednost ($t_{\text{op}} + t_{\text{lf}}$). Kao i u slučaju slabo nadziranog pristupa, za učenje modela primijenjena je logistička funkcija gubitka (5.7). Kako bi se smanjio utjecaj efekta prenaučenosti modela (engl. *overfitting*), korištena je ℓ_2 regularizacijska funkcija. Usporedba strogo nadziranog pristupa [4] u odnosu na najbolji rezultat slabo nadziranog pristupa (redak 6: 92 posto *AP*, 0.15 p_{miss}) pokazuje da predstavljeni slabo nadzirani pristup postiže bolju lokalizacijsku preciznost (za 4 posto), no veći postotak promašaja (kod strogo nadziranog pristupa samo 5 posto objekata nije lokalizirano). Jedan od mogućih razloga je složenost oba pristupa, gdje su HOG značajke izdvojene na 64 mjerila, dok su u okviru predstavljenog pristupa, SIFT značajke izdvojene na 4 mjerila (16 puta manje).

Za vrednovanje slabo nadzirane lokalizacije (retci 2 – 12), eksperimenti su organizirani u pet skupina:

- U prvoj skupini eksperimenata (retci 2 – 4) odziv se dobiva skalarnim produktom linearnog modela i Fisherova vektora okna prema izrazu (3.10).
- U drugoj i trećoj skupini (retci 5 – 8), prilikom učenja modela primijenjene su nelinearne normalizacije, a odziv okna dobiva se izravnim proračunom doprinosa pojedinog okna odzivu normaliziranog Fisherova vektora slike prema (3.12).
- U četvrtoj i petoj skupini eksperimenata (retci 11 – 12), odziv se dobiva na temelju skalarnog produkta gradijenta odziva normalizirane slike (3.22) i Fisherova vektora okna prema izrazu (3.14).

Rezultati slabo nadzirane lokalizacije ukazuju na dvije činjenice: i) primjenom rijetkih modela postiže se bolja lokalizacijska točnost u odnosu na guste modele te ii) aproksimacija odziva gradijentom (izrazi 3.22 i 3.14) rezultira nešto nižom lokalizacijskom preciznošću i značajnim ubrzanjem. U prvoj skupini eksperimenata rijetki ℓ_1 i $\ell_{2,1}$ modeli postižu bolju lokalizacijsku preciznost za do 18 postotnih bodova (redak 4 naspram retka 2) te smanjuju frekvenciju promašaja za do 6 postotnih bodova (redak 3 naspram retka 2). Primjena nelinearnih normalizacija putem izravnog proračuna doprinosa slikovnog okna (3.12) povećava lokalizacijsku preciznost

Tablica 5.2: Učinkovitost lokalizacije prometnih znakova za različite konfiguracije (M: lokalizacijski model, G: gradijent), normalizacije Fisherovih vektora i regularizacije. Uz oznaku konfiguracije navodi se broj jednadžbe prema kojoj se računa doprinos okna. Za referencu se koristi konfiguracija HOG [4], gdje je učenje obavljeno pod strogim nadzorom (prilikom učenja korištene su oznake lokacija objekata), a kao značajke se koriste histogrami orijentacije gradijenata HOG. U tom slučaju, vrijeme izvođenja uključuje t_{op} i t_{lf}.

Br.	Konf.	FV Norm.	Regularizacija	*ACD*	*AP* testiranja	p_{miss}	t_{op}/s	
1	HOG	-	l_2	-	**88**	**0.05**	10	
2	M (3.10)	-	l_2	100.0	66	0.27	8.9	
3	M (3.10)	-	l_1	6.1	81	0.21	1.9	
4	M (3.10)	-	$\ell_{2,1}$	1.1	84	0.23	0.1	
5	M (3.12)	p, ℓ_2 global	ℓ_1	3.8	85	0.16	85.3	
6	M (3.12)	p, ℓ_2 global	$\ell_{2,1}$	1.1	**92**	0.15	27.1	
7	M (3.12)	p, ℓ_2 intra	ℓ_1	5.1	76	0.25	7.6	
8	M (3.12)	p, ℓ_2 intra	$\ell_{2,1}$	0.8	88	**0.13**	0.3	
9	G (3.14)	p, ℓ_2 global	ℓ_1	3.8	82	0.17	1.2	69.4×
10	G (3.14)	p, ℓ_2 global	$\ell_{2,1}$	1.1	**87**	0.15	**0.1**	226.0×
11	G (3.14)	p, ℓ_2 intra	ℓ_1	5.1	82	0.22	1.6	4.8×
12	G (3.14)	p, ℓ_2 intra	$\ell_{2,1}$	0.8	86	**0.12**	**0.1**	4.5×

za do 8 posto te ujedno smanjuje i frekvenciju promašaja p_{miss} za 8 postotnih bodova (redak 6 naspram retka 4). Usporedba između ℓ_2 globalne metričke normalizacije (retci 5 – 6) i ℓ_2 normalizacije na razini pojedinačnih komponenata (retci 7 – 8) pokazuje da ℓ_2 intra normalizacija daje dobre rezultate isključivo kada se koristi u konjunkciji s $\ell_{2,1}$ regularizacijom. Model treniran nad FV normaliziranim po komponentama uz $\ell_{2,1}$ regularizaciju (redak 8) je nešto rjeđi u odnosu na odgovarajuću konfiguraciju koja koristi globalno normalizirane FV (redak 6) te postiže nešto lošiju preciznost lokalizacije (za 4 posto), no smanjuje frekvenciju promašaja za 2 posto. U slučaju ℓ_1 regularizacije (redak 7) postižu se lošiji rezultati u odnosu na slučaj bez ikakvih normalizacija (redak 3). Potencijalno objašnjenje utjecaja normalizacije po komponentama dano je u nastavku. Primjenom ℓ_2 intra normalizacije ujednačava se doprinos pojedinih komponenti Fisherovu vektoru, čime se smanjuje utjecaj eksplozije slikovnih riječi opisan u odjeljku 3.2. Istovremeno, ℓ_2 intra normalizacija usmjerava doprinos neuobičajenih slikovnih okana u odnosu na manji broj GMM komponenata sa značajnom vjerojatnošću mekog pridruživanja (2.23). U slučaju $\ell_{2,1}$ regularizacije to rezultira poboljšanjem lokalizacijske preciznosti, no ℓ_1 regularizacija ne uzima u obzir specifičnu strukturu Fisherovih vektora te stoga negativno utječe na lokalizacijsku preciznost.

Rezultati vrednovanja aproksimacije gradijentom u slučaju globalne metričke normalizacije (retci 9 – 10) pokazuju nešto lošiju lokalizacijsku točnost u odnosu na izravan izračun odziva okna (pad do 5 posto u odnosu na retke 5 – 6), ali i znatno ubrzanje (za dva reda veličine: redak 10 naspram retka 6). U slučaju unutar-komponentne normalizacije i $\ell_{2,1}$ regularizacije (redak 12), aproksimacija gradijentom daje najnižu vrijednost frekvencije promašaja (p_{miss} u iznosu od 0.12) u odnosu na sve slabo nadzirane konfiguracije. U slučaju ℓ_1 regularizacije, gradijent postiže bolje rezultate u odnosu na izravan proračun doprinosa okna (redak 11 naspram retka 7). Takvo se ponašanje čini neočekivanim jer aproksimacija rezultira poboljšanjem lokalizacijske točnosti u odnosu na izravan proračun. Dobiveni rezultat posljedica je činjenice da ℓ_1 regularizacija ima negativan efekt u konjunkciji s unutar-komponentnom normalizacijom. U slučaju proračuna odziva slikovnog okna primjenom gradijenta, opisani negativni efekt je manji uslijed većeg stupnja rijetkosti Fisherova vektora okna u odnosu na Fisherove vektore cjelovitih slika.

Analiza neuspjelih slučaja lokalizacije

Primjeri lokalizacije za najbolju konfiguraciju (redak 6) prema tablici 5.2 prikazani su na slici 5.2. Rezultati prikazani u gornjem retku pokazuju da metoda slabo nadzirane lokalizacije predstavljena u okviru ove disertacije uspješno pronalazi lokacije vrlo malenih objekata u urbanim prometnim scenama (prve dvije slike s lijeva na desno), ali isto tako i u ruralnim okruženjima (druge dvije slike s lijeva na desno). U donjem retku slike 5.2 prikazani su slučajevi lažnih (prve dvije slike s lijeva na desno) i neuspjelih (druge dvije slike s lijeva na desno) lokalizacija. Primjeri lažnih lokalizacija uključuju prometne znakove snimljenih sa stražnje strane i krovove kuća koji su sličnog oblika kao i prometni znakovi. U slučaju neuspjelih lokalizacija, okna najvećeg odziva identificirana su na prometnim znakovima, što pokazuje da je lokalizacijski model ispravno naučio diskriminativne slikovne riječi. Promašaji nastaju uslijed dva uzroka:

1. nedovoljne povezanosti okana ($P < 25$ posto) pa algoritam utvrđivanja lokalizacijskih poligona na temelju pojedinačnih mjerila definiran algoritmom 1 ne uspijeva generirati dovoljno velike poligone kako bi se zadovoljio uvjet $IoU > 0.5$
2. nedovoljnog broja okana ($N < 10$) pa algoritam za utvrđivanje lokalizacijskih poligona uklanja takve slučajeve iz daljnjeg razmatranja.

Posljednji slučaj promašaja najčešće se manifestira kod djelomično zaklonjenih objekata i iznimno udaljenih objekata jer je inicijalni broj slikovnih okana prisutnih na takvim objektima relativno malen.

Analiza trajanja izvođenja postupka

U okviru ovog odjeljka pobliže se razmatra utjecaj normalizacija Fisherovih vektora, aproksimacije gradijentom i različitih tipova regularizacija na vrijeme izvođenja. Valja napomenuti da su svi eksperimenti opisani u okviru odjeljka 5.2.2 provedeni na Intelovu procesoru E5-2620

Slika 5.2: Primjeri lokalizacije prometnih znakova. Gornji redak prikazuje ispravno lokalizirane objekte, dok se u donjem retku prikazuju problematični slučajevi. Dobiveni poligoni lokalizacije prikazani su žutim pravokutnicima, dok su ciljani poligoni objekata (engl. *ground-truth*) označeni crvenim pravokutnicima. Središta slikovnih okana najvećeg odziva označena su žutim točkama. Slike su prikazane u sivim tonovima (engl. *grayscale*) kako bi se naglasile lokacije okana najvećeg odziva.

frekvencije 2.00 GHz uz korištenje jedne jezgre. Shodno mjerama vremenskog profiliranja opisanim u odjeljku 5.1, dobiveni su sljedeći rezultati zajednički za sve konfiguracije u tablici 5.2:

- Vrijeme potrebno za uzorkovanje i proračun SIFT opisnika t_{lf} iznosi 0.68 sekundi, gdje se u prosjeku izdvoji $87{\cdot}10^3$ opisnika po slici.
- Vrijeme potrebno za proračun vjerojatnosti pridruživanja slikovnih okana u odnosu na GMM komponente t_{sa} iznosi 3.7 sekundi, pri čemu se izraz (2.23) evaluira za svaki od $87{\cdot}10^3$ opisnika u odnosu na K= 1024 komponente.

Usporedba u odnosu na strogo nadzirani slučaj [4] Konfiguracija koja postiže najbolji rezultat slabo nadzirane lokalizacije (redak 6 u tablici 5.2) neefikasna je u terminima vremenskog izvođenja. U tom slučaju, primjena globalne metričke normalizacije zahtjeva proračun cjelokupnog Fisherova vektora. Samim time, onemogućava se primjena trećeg koraka procesa optimizacije proračuna odziva (odjeljak 3.4.2), gdje se Fisherov vektor okna proračunava isključivo u odnosu diskriminativne komponente. Opisani negativni efekt globalne normalizacije smanjuje se primjenom normalizacije po komponentama (redak 8) i primjenom aproksimacije gradijentom (redak 10). Aproksimacija gradijentom najbolje konfiguracije (redak 10) iziskuje prosječno $t_{\mathrm{lf}} + t_{\mathrm{sa}} + t_{\mathrm{op}} = 4.48\,\mathrm{s}$, dok je lokalizacijskim modelom učenim uz strogi nadzor [4] (redak 1 u tablici 5.2) dobiveno vrijeme izvođenja $t_{\mathrm{lf}} + t_{\mathrm{op}} = 10\,\mathrm{s}$. Izravan proračun uz metričku normalizaciju po komponentama (redak 8) rezultira tek nešto većim ukupnim vremenom izvođenja u iznosu od 4.68 s.

Usporedba rijetkih i gustih modela Modeli učeni uz ℓ_1 regularizaciju postižu ubrzanje od 4.5 puta u odnosu na gusti model (redak 3 naspram retka 2). Primjena $\ell_{2,1}$ regularizacije rezultira većim stupnjem rijetkosti modela (ACD = 1.1 posto) što rezultira ubrzanjem za red veličine u odnosu na gusti model (0.1 s u odnosu 8.9 s).

Usporedba aproksimacije gradijentom u odnosu na izravan proračun odziva U slučaju globalne metričke normalizacije, aproksimacija gradijentom (retci 9 – 10) postiže značajno ubrzanje u odnosu na izravan proračun (retci 5 – 6). Za model rijedak po komponentama postiže se ubrzanje od čak dva reda veličine (redak 10 naspram retka 6). Dobiveni rezultat posljedica je opisanog negativnog efekta globalne normalizacije. Aproksimacija gradijentom također postiže ubrzanje i u slučaju normalizacije po komponentama. U odnosu na referentni rezultat dobiven izravnim proračunom odziva, bilježi se ubrzanje od gotovo 5 puta (retci 11 – 12 naspram redaka 7 – 8).

5.2.3 Vrednovanje modela temeljenih na prostornom rasporedu dijelova

U okviru ovog odjeljka naglasak je na vrednovanju reprezentacije prostornim modelima opisanim u poglavlju 4. U odnosu na model lokalizacije vrednovan u odjeljku 5.2.2, prostorni histogrami i prostorni Fisherovi vektori primjenjuju se kao opisnici druge razine za predstavljanje lokalnog rasporeda u okruženju slikovnih okana za koje lokalizacijski model prve razine daje pozitivan ishod.

Lokalizacija prostornim modelima provodi se kroz devet koraka:

1. izlučivanje SIFT opisnika gustim uzorkovanjem
2. kôdiranje SIFT opisnika u prostor Fisherovih vektora
3. sažimanje Fisherovih vektora u opisnik slike
4. primjena normalizacije potenciranjem (3.1) i ℓ_2 metričke normalizacije po komponentama (3.3) na Fisherov vektor slike
5. učenje lokalizacijskog modela $\mathbf{w_a}$
6. utvrđivanje okana pozitivnog odziva primjenom gradijenta normaliziranog odziva slike (jednadžbe 3.22 i 3.14)
7. izgradnja prostornih opisnika za predstavljanje lokalnog rasporeda u okruženju okana odabranih prethodnim korakom (algoritmi 4 i 5)
8. učenje rijetkih lokalizacijskih modela $\mathbf{w_s}$ na temelju prostornih opisnika
9. primjena lokalizacijskih modela $\mathbf{w_s}$ na razini okana i utvrđivanje lokalizacijskih poligona na temelju *svih okana pozitivnog odziva*.

Rezultati klasifikacije slika i lokalizacije objekata opisani u nastavku predstavljeni su u okviru [20].

Detalji izvedbe prostornih modela

Za izvedbu prve razine lokalizacije primjenom Fisherovih vektora i rijetkih modela koriste se postavke navedene u odjeljku 5.2.1. Prostorni opisnici definiraju se nad lokalnim okruženjima čije su dimenzije četiri puta veće u odnosu na dimenzije slikovnih okana na odgovarajućem mjerilu, konkretno koriste se dimenzije okruženja $W = H \in [64, 96, 128, 160]$ piksela. Prostorni histogrami dobivaju se diskretizacijom lokalnih okruženja u $b = 8$ odjeljaka u odnosu na obje osi. Dimenzionalnost histograma koji opisuje prostorni odnos između dvije slikovne riječi iznosi $b^2 = 64$, dok ukupna dimenzionalnost prostornog histograma lokalnog okruženja oko okna $\mathbf{x}$ iznosi $K_\mathrm{w}^2 \cdot 64$. Parametar K_w označava broj diskriminativnih komponenti modela izgleda $\mathbf{w_a}$. Za predstavljanje prostornim Fisherovim vektorima, koristi se prostorni slikovni rječnik, konkretno model mješavine Gaussovih raspodjela $\boldsymbol{\theta}_s$ s $K_\mathrm{s} = 4$ prostorne komponente dijeljen između svih parova slikovnih riječi $\boldsymbol{\theta}_s(a_1, a_2) = \ldots = \boldsymbol{\theta}_s(a_k, a_j) \ldots \boldsymbol{\theta}_s(a_{K_\mathrm{w}}, a_{K_\mathrm{w}})$. Srednja vrijednost $\boldsymbol{\mu}_s$ i varijanca $\boldsymbol{\sigma}_s$ pojedinih komponenti prostornog GMM-a odgovaraju srednjoj vrijednosti i varijanci uniformne raspodjele u četiri kvadranta kvadrata jedinične površine [108]. Dimenzionalnost Fisherova vektora za opisivanje lokalnog prostornog odnosa između dvije slikovne riječi odgovara $K_\mathrm{s} \cdot (2 \cdot D + 1) = 4 \cdot (2 \cdot 2 + 1) = 20$, a ukupna dimezionalnost prostornog Fisherova vektora lokalnog okruženja $K_\mathrm{w}^2 \cdot 20$. Prostorni Fisherovi vektori su, stoga, 3 puta kompaktniji u odnosu na reprezentaciju prostornim histogramima.

Klasifikacija slika modelom temeljenim na prostornom rasporedu dijelova

Rezultati klasifikacije slika prikazani su u tablici 5.3. U odnosu na eksperimente predstavljene u odjeljku 5.2.2, ovdje se kao mjera rijetkosti modela koristi K_w (broj diskriminativnih komponenti modela izgleda $\mathbf{w_a}$) umjesto prosječnih gustoća *ACD* i *AOD*. Primjenom mjere K_w, cilj je naglasiti fine razlike u rijetkosti među modelima. Dodatno, K_w direktno utječe na dimenzionalnost prostornih histograma i prostornih Fisherovih vektora te je, stoga, prikladnije prikazati direktno K_w u odnosu na mjeru *ACD*.

U prvoj skupini eksperimenata vrednuju se različiti tipovi modela (ℓ_2-gusti, ℓ_1-rijetki i $\ell_{2,1}$-rijetki po komponentama). Rezultati pokazuju da $\ell_{2,1}$-rijetki model (redak 3) postiže bolju klasifikacijsku točnost (*AP* testiranja) u odnosu na gusti ℓ_2 model za 7 postotnih bodova. U usporedbi sa ℓ_1-rijetkim modelom (redak 2), model rijedak po komponentama je za red veličine rjeđi (17 puta), a postiže sumjerljivu klasifikacijsku točnost na skupu za testiranje.

Dobiveni rezultati (retci 1 – 3) ukazuju na prednost $\ell_{2,1}$-rijetkih modela u odnosu na ℓ_1 i ℓ_2 regularizirane modele te se stoga u sljedećoj skupini eksperimenata (retci 4 i 5) vrednuje utjecaj nelinearnih normalizacija na modele rijetke $\ell_{2,1}$ modele. Nelinearne normalizacije poboljšavaju učinkovitost klasifikacije za 6 postotnih bodova na skupu za testiranje. Normalizacija po komponentama (redak 5) rezultira nešto rjeđim modelom u odnosu na globalnu metričku nor-

Tablica 5.3: Učinkovitost klasifikacije slika na skupu prometnih znakova za model temeljen na prostornom rasporedu dijelova. Razmatraju se različite konfiguracije (M: klasifikacijski model prve razine temeljen na značajkama izgleda $\mathbf{w_a}$, SH: prostorni klasifikacijski model temeljen na prostornim histogramima, SFV: prostorni klasifikacijski model temeljen na prostornim Fisherovim vektora), normalizacije (p: normalizacija potenciranjem, ℓ_2 global: metrička normalizacija cjelokupnog FV, ℓ_2 intra: metrička normalizacija po komponentama), te regularizacije (ℓ_2, ℓ_1 i $\ell_{2,1}$). Parametar K_w označava broj komponenti modela izgleda $\mathbf{w_a}$ sa normom različitom od ništice.

Br.	Konfiguracija	FV normalizacija	Regularizacija	K_w	*AP* učenja	*AP* testiranja
1	M	-	ℓ_2	1024	100	64.0
2	M [158]	-	ℓ_1	185	98	71.9
3	M	-	$\ell_{2,1}$	11	80	71.1
4	M	p, ℓ_2 global	$\ell_{2,1}$	10	83	76.9
5	M	p, ℓ_2 intra	$\ell_{2,1}$	7	81	76.8
6	SH	p, ℓ_2 intra	$\ell_{2,1}$	7	92	**81.8**
7	SFV	p, ℓ_2 intra	$\ell_{2,1}$	7	94	81.2

malizaciju (redak 4), bez gubitka klasifikacijske točnosti (7 naspram 10 komponenata).

U posljednjoj skupini eksperimenata (retci 6 i 7), vrednuje se klasifikacija prostornim modelima. Prostorni se opisnici slika grade nad oknima za koje model prve razine daje pozitivan ishod. Zatim se dobiveni opisnici lokalnih okruženja sažimaju na razini slike i obavlja se skalarni produkt prostornog opisnika slike i naučenog prostornog klasifikacijskog modela. Rezultati pokazuju da modeli prostornog rasporeda poboljšavaju klasifikacijsku točnost na skupu za testiranje za do 4 postotna boda u odnosu na rezultat koji se temelji isključivo na značajkama izgleda (retci 4 i 5). Rijedak $\ell_{2,1}$ model izdvaja ukupno $K_\mathrm{w} = 7$ diskriminativnih komponenata, odnosno prostornim se opisnicima modelira raspored između $K_\mathrm{w}^2 = 49$ parova slikovnih riječi. Prostornih histogrami SH i prostorni Fisherovi vektori SFV daju sumjerljive rezultate, no prostorni su Fisherovi vektori 3 puta kompaktniji, gdje dimenzionalnost SH iznosi $K_\mathrm{w}^2 \cdot b^2 = 49 \cdot 64 = 3136$, dok dimenzionalnost SFV iznosi $K_\mathrm{w}^2 \cdot K_\mathrm{s} \cdot (2D+1) = 49 \cdot 4 \cdot 5 = 980$. Veća kompaktnost SFV opisnika rezultira i manjom dimenzionalnošću klasifikacijskog modela, odnosno većom efikasnošću izvođenja.

Lokalizacija objekata prostornim modelom

Kvalitativni rezultat lokalizacije prometnih znakova prikazani su tablicom 5.4.

Lokalizacijski poligoni za konfiguracije temeljene isključivo na značajkama izgleda (retci 1 – 4) dobiveni su primjenom algoritma 1 nad T=100 okana najvećeg odziva. Rezultati ukazuju na značaj nelinearnih normalizacija za zadatak lokalizacije, gdje normalizacije povećavaju

Tablica 5.4: Učinkovitost lokalizacije prometnih znakova s naglaskom na prostorni model. Parametar T označava broj okana najvišeg odziva koja su korištena za proračun lokalizacijskog poligona. Parametar K_w označava broj odabranih komponenti modela $\mathbf{w_a}$, dok p_{miss} označava frekvenciju promašaja na krajnje desnoj točki *PR* krivulje.

Br.	Konfiguracija	FV normalizacija	Regularizacija	K_w	T	*AP* testiranja	p_{miss}
1	M [158]	-	l_1	64	100	72 .0	0.13
2	M	-	$\ell_{2,1}$	11	100	74.0	0.25
3	M	p, ℓ_2 intra	$\ell_{2,1}$	7	100	77.4	0.11
4	G	p, ℓ_2 intra	$\ell_{2,1}$	7	100	77.0	0.16
5	G + SH	p, ℓ_2 intra	$\ell_{2,1}$	7	sva	75.0	0.14
6	G + SFV	p, ℓ_2 intra	$\ell_{2,1}$	7	sva	**81.0**	**0.11**

preciznost lokalizacije za 3 postotna boda, a ujedno smanjuju i frekvenciju promašaja na 11 posto (redak 3 naspram retka 2). Aproksimacija gradijentom (redak 4) ne rezultira gubitkom lokalizacijske preciznosti, no povećava frekvenciju promašaja za 5 postotnih bodova.

Lokalizacija prostornim modelima prikazana je retcima 5 i 6. Prostorni su opisnici izgrađeni nad rezultatima dobivenim primjenom gradijenta normaliziranog odziva slike (3.22). Ovdje su lokalizacijski poligoni dobiveni na temelju svih opisnika lokalnih okruženja za koje prostorni model $\mathbf{w_s}$ daje pozitivan ishod. Prostorni histogrami SH (redak 5) daju nešto lošiju lokalizacijsku preciznost u odnosu na odgovarajući model izgleda (redak 4), no umanjuju frekvenciju promašaja za 2 postotna boda i općenito gledajući, smanjuju broj parametara algoritma (uklanjaju potrebu za parametrom T jer se koriste sva okna pozitivnog ishoda). Prostorni Fisherovi vektori SFV (redak 6) izgrađeni nad odzivima gradijenta daju najbolji rezultat lokalizacije (*AP* testiranja = 81 posto, $p_{\text{miss}} = 0.11$). Točnije, povećavaju učinkovitost lokalizacije za 4 postotna boda u odnosu na model prve razine lokalizacije (retci 3 i 4) te smanjuju frekvenciju promašaja za 5 postotnih bodova u odnosu na aproksimaciju gradijentom (redak 4).

Primjeri lokalizacija prostornim Fisherovim vektorima SFV ilustrirani su slikom 5.3. Rezultati pokazuju da se modeliranjem lokalnog prostornog rasporeda uklanjaju lažni pozitivi koji se pojavljuju u okviru teških pozadina. Metoda uspješno lokalizira udaljene objekte i u stanju je lokalizirati i diferencirati između udaljenih vrlo bliskih objekata (primjer broj 2 s lijeva na desno u slici 5.3). Druge dvije slike daju primjere lažnih odziva, gdje se lažni odzivi odnose na višestruke poligone lokalizacija na samim objektima. Dobiveni lažni odzivi posljedica su činjenice da je u okviru algoritma 3 za uklanjanje višestrukih poligona lokalizacija primijenjen samo prvi korak koji za kriterij uklanjanja uzima isključivo veličinu odziva slikovnog okna.

Slika 5.3: Rezultati lokalizacije prostornim Fisherovim vektorima SFV. Prve dvije slike s lijeva na desno ilustriraju primjere uspješnih lokalizacija vrlo malenih objekata. Druge dvije slike ilustriraju primjere lažnih odziva. Okna pozitivnog odziva prikazana su različitim bojama, gdje boja okna odgovara pripadnosti diskriminativnoj slikovnoj riječi.

Vremenska učinkovitost prostornih modela

Primjena druge razine lokalizacije uključuje proračun prostornih opisnika i primjenu prostornog modela $\mathbf{w_s}$ na razini okna. Opisani proces traje oko 0.2 s u slučaju prostornih histograma SH i prostornih Fisherovih vektora SFV. U usporedbi s prostornim histogramima, prostorni Fisherovi vektori uključuju proračun gradijenata logaritamske izglednosti u odnosu na parametre prostornog slikovnog rječnika $\boldsymbol{\theta}_s$. Primjenom optimizacija proračuna prostornih Fisherovih vektora opisanih u odjeljku 4.2.1, ti se gradijenti mogu unaprijed izračunati i pohraniti u priručnu tablicu (engl. *look-up table*). U fazi lokalizacije, s obzirom na izračunate prostorne vektore odmaka između okana koja odgovaraju različitim slikovnim riječima, dohvaćaju se odgovarajući Fisherovi vektori u odnosu na prostorni slikovni rječnik iz priručne tablice pohranjene u glavnoj memoriji računala. S obzirom da je prostorni GMM dijeljen između svih K_w^2 parova slikovnih riječi, priručna tablica za prostorne Fisherove vektore zauzima od 16.3 KB (za okruženja dimenzija 64×64 piksela) do 102.4 KB (za okruženja dimenzija 160×160 piksela) radne memorije. Za lokalno okruženje pozitivno označenog okna $\mathbf{x}$ dimenzija $W \times H$ može se izdvojiti upravo $W \times H$ parova vektora odmaka $d(\mathbf{x},\mathbf{z}) \in \mathbb{R}^2$. S obzirom na prostorni GMM s K_s komponenti, dimenzionalnost Fisherova vektora za par slikovnih riječi iznosi $K_s \cdot (2 \cdot D + 1) = 4 \cdot (2 \cdot 2 + 1) = 20$. Fisherovi se vektori pohranjuju kao vrijednosti sa posmičnim zarezom, tako da pojedini Fisherov vektor zauzima 4 B. Za lokalno okruženje dimenzija $W = H = 64$, svi Fisherovi vektori iziskuju $W \cdot H \cdot 4 = 16384$ B. Analogan proračun vrijedi i za ostale veličine lokalnih okruženja. S obzirom na količine radne memorije u današnjim računalima, veličine priručnih tablica gotovo su zanemarive. Zaključno, primjenom prostornih Fisherovih vektora postiže se bolja lokalizacijska točnost uz zadržavanje brzine izvođenja.

5.3 Lokalizacija pješačkih prijelaza

U okviru ovog odjeljka razmatra se postupak učenja lokalizacijskog modela na temelju geopodataka prikupljenih radom mnoštva (engl. *crowdsourced data*) i georeferenciranih video zapisa

u svrhu automatizacije kartiranja pješačkih prijelaza. U nastavku odjeljka opisan je postupak stvaranja slabo označenog skupa slika, a zatim su dani rezultati klasifikacije slika i lokalizacije objekata. Razmatra se analiza vremenskog izvođenja i neuspjelih lokalizacija te utjecaj *IoU* praga na lokalizacijsku preciznost.

5.3.1 Prikupljanje slabo označenog skupa slika primjenom dobrovoljno prikupljenih geopodataka

U okviru ovog odjeljka opisan je postupak stvaranja slabo označenog skupa slika na temelju OpenStreetMap karte [159] i georeferenciranih video zapisa [70].

Detaljno se opisuje struktura OpenStreetMap podataka te struktura i načini prikupljanja georeferenciranih video zapisa. Postupak uparivanja OSM oznaka u odnosu na georeferencirani video zapis formalno je opisan u obliku algoritma 6. Dane su konkretne statistike skupa podataka dobivenog primjenom opisanog algoritma za OSM značajke označene s „highway = crossing", odnosno za prikupljanje slika pješačkih prijelaza.

Struktura OSM geo-podataka

OpenStreetMap geoprostorna baza podataka sačinjena je od tri osnovna elementa kojima se opisuju različiti tipovi geopodataka i odnosi među njima [160]:

1. čvorova (engl. *nodes*)
2. putova (engl. *ways*)
3. relacija (engl. *relations*).

Čvorovi čine temeljnu jedinicu OSM geoprostorne baze podataka. Predstavljeni su GPS koordinatama u obliku WGS 84 [161] geografske širine (engl. *latitude*) i dužine (engl. *longitude*).

Na temelju čvorova, definiraju se putovi kao uređene lista čvorova. Putovi služe za predstavljanje i) linijskih objekata kao što su, primjerice, ceste ili rijeke te ii) poligona kojima se modeliraju različite građevine, parkovi, jezera i mnoge druge geografske značajke.

Relacije su definirane kao uređene liste čvorova, putova ili čak drugih relacija, a služe za opisivanje logičkih ili geografskih odnosa između pojedinih elemenata. Primjeri logičkih relacija uključuju zabrane skretanja (engl. *turn restrictions*) na prometnim križanjima ili zabrane prometa za teretna vozila uvjetovane prisutnošću odgovarajućih prometnih znakova. Primjeri geografskih relacija uključuju autobusne ili tramvajske linije sačinjene od putova koji predstavljaju trase putovanja odgovarajućeg vozila i čvorova koji predstavljaju stajališta. Drugi primjeri geografskih relacija uključuju višestruke poligone (engl. *multipolygons*) šuma ili jezera koji mogu biti sačinjeni od više pojedinačnih poligona ili to mogu biti poligoni s rupama.

Svakome od navedenih elemenata pridjeljuje se semantičko značenje putem oznaka (engl. *tags*) definiranih prema načelu „ključ = vrijednost". Uloga ključeva jest organizirati elemente u raz-

rede (kategorije). Tako su, primjerice, objekti cestovne infrastrukture označeni ključem „highway" koji poprima širok spektar vrijednosti. Neki od primjera uključuju:

- „highway" = „traffic_signals" za označavanje regulacije prometa svjetlosnim znakovima (semaforima) na danim lokacijama
- „highway" = „give_way" za označavanje prometnog znaka izričitih naredbi „križanje s cestom s prednošću prolaska"
- „highway" = „crossing" za označavanje pješačkih prijelaza.

OpenStreetMap podacima može se pristupiti preko različitih izvora [162], [62] koji omogućuju preuzimanje podataka za pojedine gradove, regije ili države, ili putem servisa [163] koji omogućuje pretraživanje i preuzimanje specifičnih geografskih značajki. Primjenom odgovarajućih upita preko servisa [163] mogu se preuzeti svi OSM elementi koji označavaju pješačke prijelaze („highway" = „crossing") na nekom području. Slika 5.4 ilustrira primjer takvog upita. Primjenom takvog upita dohvaćene su lokacije pješačkih prijelaza za područja gradova Siska i Karlovca. Dobivene lokacije korištene su za stvaranje slabo označenog skupa podataka pješačkih prijelaza.

Georeferencirani video

Georeferencirani video zapis uključuje:

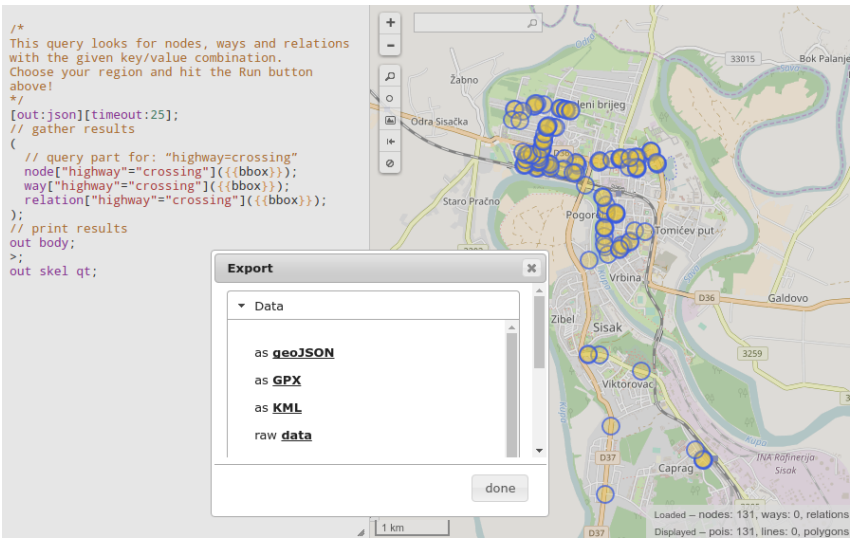
- datoteku videa u nekom od standardnih formata (primjerice, mp4, m4v, 3gp i drugi)
- tekstualnu datoteku koja sadrži prostorno-vremenske podatke vezane uz pojedini okvir (engl. *frame*) kao što su vrijeme, GPS pozicija ili smjer kamere kojom je zapis prikupljen.

Na tržištu se mogu pronaći različita tehnološka rješenja [159] za prikupljanje georeferenciranog videa od kojih neki primjeri uključuju kamere i kamkordere opremljene GPS prijemnicima, ili primjenske programe za pametne telefone kao što su Mapillary [67] ili OpenStreetView [66].

Za potrebe ove disertacije, korišteni su video zapisi prikupljeni u okviru projekta E-ceste za gradove Karlovac i Sisak [70]. Na području navedenih gradova, prikupljeno je ukupno 1536 video zapisa primjenom uređaja različite kvalitete kao što je, primjerice, video kamera *GoPro Hero 2* koju odlikuje visoka rezolucija snimanja (1080p) i oštrina slike [164] ili pak kamere niže rezolucije (720p). Video zapisi pokrivaju urbane prometne scene (80 posto video materijala), ali i ruralne prometne scene (20 posto video materijala). Ukupna površina pokrivena navedenim video materijalima iznosi $3.5\,\mathrm{km}^2$ za područje grada Karlovca te $8.78\,\mathrm{km}^2$ za područje grada Siska.

Svakoj video datoteci dodijeljena je tekstualna datoteka u JSON formatu. Datoteka sadrži niz prostorno-vremenskih objekata od kojih je svaki opisan:

- GPS lokacijom u obliku "coordinates": [<zemljopisna_dužina>, <zemljopisna_širina>]
- vremenskim odmakom u sekundama odnosu na početak video zapisa u obliku "time": <vremenski_odmak>.



Slika 5.4: Izdvajanje i preuzimanje podataka za značajke označene s „highway" = „crossing" na razini grada Siska putem servisa [163]. S lijeve strane prikazan je upit kojim se dohvaćaju objekti, dok su s desne strane prikazane lokacije pješačkih prijelaza u okviru OpenStreeMap karte. U središtu slike prikazano je sučelje za preuzimanje podataka u različitim formatima.

Konkretan primjer prostorno-vremenskih objekata prikazan je na slici 5.5, gdje pojedini objekt poprima sljedeći oblik {"coordinates": [15.603174, 45.478279], "time": 53.5 }.

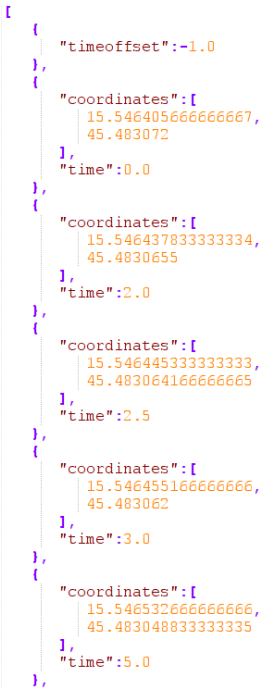
Uparivanje OSM objekata u odnosu na georeferencirani video

Prije samog procesa povezivanja lokacija OSM objekata s georeferenciranim videom s ciljem izdvajanja slika tog objekta iz videa, potrebna je obrada prostorno-vremenskih podataka vezanih uz video. Niz geoprostornih podataka [{"coordinates": [lon_0,lat_0], "time": t_0 }, ..., {"coordinates": [lon_i,lat_i], "time": t_i }, ..., {"coordinates": [lon_n,lat_n], "time": t_n }] dijeli se u niz segmenata $\mathbf{G} = \{\mathbf{g}_j\}$:, gdje duljina pojedinog segmenta iznosi barem 5 m. Dobiveni se segmenti koriste kako bi se na brz i jednostavan način odredili dijelovi video zapisa koji potencijalno sadrže snimke objekata.

Postupak uparivanja prikazan je algoritmom 6, dok slika 5.6 ilustrira primjer uparivanja za konkretan OSM čvor osm_id = 2043645281. Ulaz u algoritam uparivanja čine OSM čvor definiran odgovarajućim GPS koordinatama **n**=(lat, lon) te skup video zapisa $\mathbf{V} = (\mathbf{F}, \mathbf{G})$ koji obuhvaća: i) skup video okvira (engl. *frames*) i pripadajućih GPS lokacija $\mathbf{F} = \{\mathbf{I}_i, \mathbf{p}_i\}$ te ii) skup segmenata $\mathbf{G} = \{\mathbf{g}_j\}$ duljine barem 5 m. Algoritam je parametriziran sljedećim vrijednostima:

- T - označava broj slika koje se izdvajaju za određeni fizički OSM čvor
- Δd - označava udaljenost GPS lokacija na kojima se izdvajaju uzastopne slike OSM objekta.

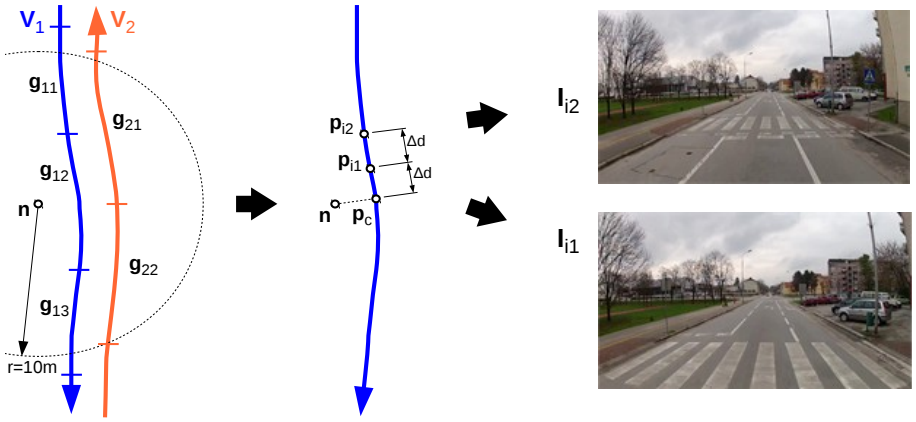
Za svaki OSM čvor **n**, algoritam razmatra sve segmente $\mathbf{g}_j$ koji: i) se nalaze u radijusu od 10 m u odnosu na čvor **n** te ii) zadovoljavaju uvjet da su bliži čvoru **n** u odnosu na susjedne segmente



```
[
  {
    "timeoffset":-1.0
  },
  {
    "coordinates":[
      15.546405666666667,
      45.483072
    ],
    "time":0.0
  },
  {
    "coordinates":[
      15.546437833333334,
      45.4830655
    ],
    "time":2.0
  },
  {
    "coordinates":[
      15.546445333333333,
      45.483064166666665
    ],
    "time":2.5
  },
  {
    "coordinates":[
      15.546455166666666,
      45.483062
    ],
    "time":3.0
  },
  {
    "coordinates":[
      15.546532666666666,
      45.483048833333335
    ],
    "time":5.0
  },
```


Slika 5.5: Fragment prostorno-vremenske datoteke vezane uz georeferencirani video. Datoteku sačinjava niz objekata od kojih je svaki opisan prostornim koordinatama (niz *coordinates*) i vremenskim odmakom u odnosu na početak videa (*time*).

$\mathbf{g}_{j-1}$ i $\mathbf{g}_{j+1}$. Osnovna ideja jest pronaći na svakom takvom segmentu $\mathbf{g}_j$ najbližu lokaciju u odnosu na $\mathbf{n}$ označenu sa $\mathbf{p}_c$ i počevši od točke $\mathbf{p}_c$ prema unazad načiniti T snimaka čvora $\mathbf{n}$ udaljenih Δd metara.



Slika 5.6: Primjer uparivanja OSM čvora osm_id = 2043645281, koji se nalazi na $\mathbf{n} = 45.487347$ N, 15.556345 E u odnosu na georeferencirane video zapise označene sa $\mathbf{V}_1$ i $\mathbf{V}_2$. U prvom koraku, algoritam uparivanja pronalazi segmente $\mathbf{g_{12}}$ i $\mathbf{g_{21}}$. Detalji uparivanja dani su za segment $\mathbf{g_{12}}$, a analogno vrijede i za segment $\mathbf{g_{21}}$. Najprije se određuju GPS koordinate točke $\mathbf{p}_c$, najbliže u odnosu na čvor $\mathbf{n}$. Zatim se unatražno u odnosu na lokaciju $\mathbf{p}_c$ izdvajaju $T = 2$ slike uz pomak od Δd=3 m.

Algoritam 6 Uparivanje OSM objekata u odnosu na georeferencirani video

Parametri:

T: željeni broj slika po OSM čvoru,

Δd: udaljenost između pojedinih slika;

Ulaz:

OSM čvor **n**=(lat, lon),

Video zapisi $\mathbf{V} = (\mathbf{F}, \mathbf{G})$ gdje

i) $\mathbf{F} = \{\mathbf{I}_i, \mathbf{p}_i\}$: skup video okvira i pripadajućih GPS lokacija

ii) $\mathbf{G} = \{\mathbf{g}_j\}$: skup segmenta duljine ≥ 5 m

(gdje $\mathbf{g}_j$ predstavlja niz GPS lokacija $\{\mathbf{p}_k\}, k \in \mathbf{g}_j$)

Za sve video zapise $\mathbf{V} = (\mathbf{F}, \mathbf{G})$

  Za sve $\mathbf{g}_j \in \mathbf{G} : \|\mathbf{g}_j - \mathbf{n}\| < \min(\|\mathbf{g}_{j-1} - \mathbf{n}\|, \|\mathbf{g}_{j+1} - \mathbf{n}\|)$

    Pronađi $c = \arg\min_{k \in \mathbf{g}_j} \|\mathbf{p}_k - \mathbf{n}\|$,

    Ako $\|\mathbf{p}_c - \mathbf{n}\| \leq 10$ m

      Za t $\in 1, 2, \ldots T$

        Pronađi $i_t = \arg\min_i (t \cdot \Delta d - \sum_{k=i+1}^{c} \|\mathbf{p}_k - \mathbf{p}_{k-1}\|)^2$

        Ako $i_t = 0$

          Prekini

        Izdvoji sliku $\mathbf{I}_{i_t}$

Izlaz: Izdvojene slike označene kao „objekti”

Odabir parametara T i Δd Cilj lokalizacije u okviru automatizacije digitalne kartografije jest pronaći precizne lokacije objekata. Lokalizacija malenih objekata udaljenih u odnosu na

kameru nije primarni zadatak jer se za njih ne može precizno odrediti GPS pozicija. Iz navedenih je razloga potrebno odabrati parametre T i Δd na način da su izdvojene slike u neposrednoj blizini OSM objekta. Istovremeno, potrebno je razviti robustan algoritam neosjetljiv na razlike između OSM geometrije i GPS podataka prikupljenih u okviru georeferenciranog videa. U eksperimentima se, stoga, za svaki čvor i pojedini segment u radijusu od 10 m izdvajaju $T = 3$ slike svakih $\Delta d = 3$ m, no naglasak je na lokalizaciji objekata bliskih očištu.

Otpornost na šum u podacima Zbog specifične prirode problema (OSM lokacije objekata i georeferencirani video zapisi prikupljeni su od strane različitih korisnika i GPS uređaja), opisanim algoritmom će se uz valjane slike objekata prikupiti i slike koje ne sadrže objekte traženog razreda. U procesu stvaranja slabo označenog skupa podataka opisanim algoritmom, detektirano je 15 posto takvih slika koje su uklonjene iz skupa pozitiva ručnim probirom. Zanimljiv smjer budućeg rada uključivao bi učenje lokalizacijskog modela nad skupom slika na kojem nije obavljeno filtriranje, odnosno potpuna automatizacija digitalnog kartiranja.

Statistike skupa podataka pješačkih prijelaza

Opisanim algoritmom uparivanja, za svaki je OSM pješački prijelaz izdvojeno u prosjeku šest slika iz različitih video zapisa i različitih udaljenosti u odnosu na objekt. Dobivene slike snimljene su iz različitih kutova gledišta (primjerice, sprijeda, straga, s lijeva ili s desna) i različitih osvjetljenja uvjetovanih vremenskim uvjetima (oblačno ili sunčano vrijeme). Gornji redak slike 5.7 ilustrira unutar-razrednu varijabilnost prikupljenog skupa podataka za razred pješačkih prijelaza. Algoritmom 6 izdvojeno je ukupno 1259 slika podijeljenih ručnim probirom u 1067 pozitiva i 192 negativa (slika bez pješačkih prijelaza). Skup negativa proširen je sa 1122 slike izdvojene slučajnim odabirom iz dostupnih video zapisa te naknadnom ručnom verifikacijom da izdvojene slike uistinu ne sadrže pješačke prijelaze. Cjelokupni proces generiranja slabo označenog skupa podataka obavljen je u trajanju od 1 sat i 50 minuta, pri čemu probir OSM slika zahtjeva 20 minuta, dok izdvajanje negativnih slika iziskuje 1.5 sati. Donji redak slike 5.7 ilustrira reprezentativne primjere prikupljenih negativa. Budući da je 80 posto video materijala prikupljeno u urbanim sredinama, negativne slike sadrže mnoštvo uzoraka koji se također pojavljuju kao elementi pješačkih prijelaza.

Prikupljeni skup podataka podijeljen je u dva disjunktna skupa: skup za učenje s 1299 slike i skup za testiranje s 1082 slika. Prilikom raspodjele slika u skupove za treniranje i testiranje, posebna je pažnja usmjerena na činjenicu da sve slike određenog fizičkog prijelaza budu dodijeljene jednakom podskupu. Kako bi se omogućilo vrednovanje lokalizacijske točnosti, podskup za testiranje označen je aproksimacijama poligona pješačkih prijelaza. Na 484 pozitivne slike u skupu za testiranje, označeno je 674 poligona pješačkih prijelaza pri čemu 337 slika sadrži po jedan objekt, 120 sadrži dva objekta, 14 slika sadrži tri objekta i dodatnih 13 slika sadrži više od

Slika 5.7: Slabo označeni skup podataka za pješačke prijelaze: primjeri slika prikupljenih uparivanjem lokacija OSM čvorova označenih sa ”highway” = ”crossing” na georeferencirani video. Gornji redak: slike OSM objekta (osm_id 981409265 pozicioniranog na 45.483031 N, 15.546749 E) izdvojene iz različitih video zapisa. S lijeva na desno: slika svježe obojenog pješačkog prijelaza, slika snimljena kamerom montiranom na bicikl, slika izbijeljenog pješačkog prijelaza snimljena iz neposredne blizine te slika djelomično zaklonjenog objekta snimljena sa udaljenosti od 20-tak metara u odnosu na objekt. Donji redak: primjeri negativnih slika. Negativne slike sadrže mnoštvo objekata sa sličnim uzorcima koji se mogu pronaći i na pješačkim prijelazima (primjerice pješački otoci, parkirališne površine, zaštitne ograde ili autobusna stajališta).

tri objekta. Neke od slika sadrže nekoliko objekta udaljenih od točke očišta. Kompletnosti radi, označeni su i takvi objekti (iako se za njih ne može precizno odrediti GPS pozicija). Zaključno, 30 posto označenih objekata manje je od 1 posto površine slike, 48 posto je između 2 i 7 posto, a preostalih 22 posto veće je od 7 posto površine slike.

5.3.2 Detalji izvedbe

Za potrebe lokalizacije pješačkih prijelaza kao lokalni opisnici koriste se konvolucijske značajke izdvojene iz posljednjeg konvolucijskog sloja duboke mreže VGG-E [39] (conv3-512) prikazane na slici 2.5 u okviru odjeljka 2.1.2. Duboka konvolucijska mreža VGG-E učena je na skupu podataka ImageNet [93] i kao takva primijenjena na slike dobivene primjenom OSM podataka prikupljenih radom mnoštva. Lokalne su značajke izdvojene na tri mjerila skaliranjem slike za faktor 2^m, gdje mjerilo $m \in \{0, -0.5, -1\}$. Dimenzionalnost izdvojenih značajki odgovara broju neurona (kanala) u petom sloju VGG-E mreže $D = 512$. Za potrebe slikovnog rječnika, koristi se model mješavine Gaussovih raspodjela s $K = 128$ Gaussovih komponenata s dijagonalnim matricama kovarijacije. GMM je učen na slučajno odabranom uzorku od $2 \cdot 10^6$ lokalnih opisnika primjenom algoritma optimizacije očekivanja implementiranog u okviru programske biblioteke *Yael* [155]. Kôdiranjem lokalnih opisnika dobivaju se Fisherovi vektori dimenzija $K \cdot (2D+1) = 128 \cdot (2 \cdot 512 + 1) = 131200$.

Lokalizacijski model učen je na temelju dobivenih Fisherovih vektora slika i oznaka prisutnosti objekata u slikama minimizacijom logističkog gubitka (5.7). Za minimizaciju funkcije gubitka koristi se algoritam *FISTA* (engl. *Fast Iterative Shrinkage-Thresholding Algorithm*) [165] izveden u okviru programske biblioteke *SPAMS* [156]. Kao i u slučaju prometnih znakova,

Tablica 5.5: Učinkovitost klasifikacije slika pješačkih prijelaza u odnosu na različite normalizacije Fisherovih vektora (p: potenciranje, ℓ_2 globalna metrička, ℓ_2 metrička unutar komponente) i regularizacije (ℓ_1, ℓ_2, $\ell_{2,1}$: ℓ_2 unutar komponente, ℓ_1 između komponenata). Prosječna cjelokupna gustoća (engl. *average overall density, AOD*) označava udio koeficijenata modela različitih od ništice, dok prosječna gustoća komponenata (engl. *average component density, ACD*) označava udio komponenti modela sa normom različitom od ništice.

Br.	FV normalizacija	Regularizacija	*AOD*	*ACD*	*AP* učenja	*AP* testiranja
1	-	ℓ_2	77.4	100	97	95
2	-	ℓ_1	2.1	78.9	94	95
3	-	$\ell_{2,1}$	8.5	14.8	95	95
4	p, ℓ_2 global	ℓ_2	77.4	100	100	97
5	p, ℓ_2 global	ℓ_1	0.2	30.5	100	98
6	p, ℓ_2 global	$\ell_{2,1}$	3.8	7.0	100	98
7	p, ℓ_2 intra	ℓ_2	77.4	100	100	97
8	p, ℓ_2 intra	ℓ_1	0.1	21.1	97	98
9	p, ℓ_2 intra	$\ell_{2,1}$	2.6	**4.7**	100	**98**

vrednuju se različite regularizacijske funkcije opisane u odjeljku 3.3.1, pri čemu je parametar λ određen na temelju unakrsne provjere s 10 odjeljaka (engl. *10-fold cross-validation*).

Lokalizacijski poligoni dobivaju se na temelju okana pozitivnog odziva primjenom pristupa temeljenog na ujedinjenim mapama odziva preko više mjerila. Navedeni pristup detaljno je opisan u odjeljku 3.5.2, a odabran je jer su preliminarni eksperimenti s pristupom temeljenim na pojedinačnom mjerilu davali loše rezultate (generirali su nedovoljno velike poligone lokalizacije). Učinkovitost lokalizacije vrednuje se prema kriteriju 5.3 [131], pri čemu se razmatraju različite vrijednosti *IoU* praga, gdje $IoU \in \{0.1, 0.2, 0.3, 0.4, 0.5\}$.

5.3.3 Rezultati klasifikacije slika pješačkih prijelaza

Rezultati klasifikacije slika prikazani su u okviru tablice 5.5. Prilikom prezentacije rezultata primjenjuje se organizacija eksperimenata u tri skupine kao u i slučaju prometnih znakova u odjeljku 5.2.2.

Rezultati pokazuju da rijetki i gusti modeli rezultiraju sumjerljivom klasifikacijskom točnošću za sve konfiguracije, neovisno o primjeni normalizacija ili uporabi gradijenta, odnosno modela. Prednost rijetkih modela jest da postižu jednako dobre rezultate koristeći samo dio cjelokupne visokodimenzionalne reprezentacije slike.

U prvoj se skupini vrednuju modeli učeni nad Fisherovim vektorima bez normalizacija (retci 1 – 3). Model rijedak po komponentama ($\ell_{2,1}$ regularizacijska funkcija, redak 3) pokazuje

najveći stupanj uspješnosti prilikom probira diskriminativnih slikovnih riječi. U odnosu na ℓ_1 rijedak model koji koristi gotovo *ACD* = 80 posto komponenti modela, $\ell_{2,1}$ model koristi pet puta manje komponenti modela (14.8 naspram 78.9).

U drugoj skupini eksperimenata (retci 4 – 6) vrednuje se doprinos nelinearnih normalizacija. Rezultati ukazuju da primjena normalizacija:

1. povećava klasifikacijsku točnost za do tri postotna boda te istovremeno
2. povećava stupanj rijetkosti za ℓ_1 i $\ell_{2,1}$ rijetke modele.

Normalizacijom se smanjuje raspon vrijednosti Fisherovih vektora što utječe na to da model donosi odluku provjerom prisutnosti određene slikovne riječi u odnosu na njezinu učestalost u slici.

Usporedba normalizacije po komponentama u odnosu na globalnu metričku normalizaciju (retci 7 – 9 naspram redaka 4 – 6) pokazuje da obje normalizacije postižu istovjetnu učinkovitost klasifikacije na skupu za testiranje. Unutar-komponentna normalizacija posebice daje dobre rezultate u konjunkciji s $\ell_{2,1}$ rijetkim modelom (redak 9), gdje se postiže najbolji omjer gustoće modela i klasifikacijske učinkovitosti (*ACD* iznosi tek 4.7 posto, dok je *AP* testiranja 98 posto). Dobiveno ponašanje rezultat je činjenice da normalizacija po komponentama i $\ell_{2,1}$ regularizacija uzimaju u obzir specifičnu blokovsku strukturu Fisherovih vektora.

5.3.4 Rezultati lokalizacije pješačkih prijelaza

Kvalitativni rezultati lokalizacije pješačkih prijelaza prikazani su tablicom 5.6. Slika 5.8 ilustrira primjere lokalizacijskih poligona označene žutom bojom za različite konfiguracije iz tablice 5.6.

S obzirom na polu-automatski način prikupljanja slika s pješačkim prijelazima i veliku unutar-razrednu varijabilnost (ilustriranu na slici 5.7), za verifikaciju učinkovitosti lokalizacije koristi se *IoU* [131] prag od 0.10 prema izrazu (5.3). U odjeljku 5.3.6 detaljno se razmatra i) utjecaj *IoU* praga na lokalizacijsku točnost te ii) korelacija veličine objekta u odnosu na prosječnu vrijednost izraza (5.3).

Rezultati prve skupine eksperimenata pokazuju da rijetki modeli daju bolju lokalizacijsku točnost u odnosu na gusti ℓ_2 model: prosječna preciznost (*AP* testiranja) u rijetkih modela je veća za 16 postotnih bodova, a ujedno je i frekvencija promašaja p_{miss} manja za 15 postotnih bodova. Prvi redak u slici 5.8 također potvrđuje dobivene rezultate, gdje ℓ_2 model (krajnje lijevo) odabire značajan broj piksela u pozadini te nije u stanju prepoznati dva objekta u slici, već za oba generira jedinstveni lokalizacijski poligon. Model učen uz primjenu $\ell_{2,1}$ regularizacije (redak 3) postiže nešto bolju lokalizacijsku preciznost u odnosu na ℓ_1-rijedak model unatoč činjenici da koristi znatno manji broj slikovnih riječi. Slika 5.8 pokazuje kako $\ell_{2,1}$ model (krajnje desno u prvom retku) odabire znatno manje piksela u pozadini u odnosu na ℓ_1 model.

Primjenom normalizacije potenciranjem (oznaka *p* u trećem stupcu tablice 5.6) i globalne

Tablica 5.6: Učinkovitost lokalizacije pješačkih prijelaza za različite konfiguracije (M: lokalizacijski model, G: gradijent), normalizacije Fisherovih vektora i regularizacije. Uz oznaku konfiguracije navodi se broj jednadžbe prema kojoj se računa doprinos okna. Mjera p_{miss} označava frekvenciju promašaja u krajnje desnoj točki *PR* krivulje. Mjera t_{op} označava prosječno vrijeme potrebno za izračun odziva okana u slici. Za konfiguracije koje uključuju gradijent (retci 8 – 11), krajnje desno se prikazuje ubrzanje u odnosu na izravan proračun odziva okna (retci 4 – 7).

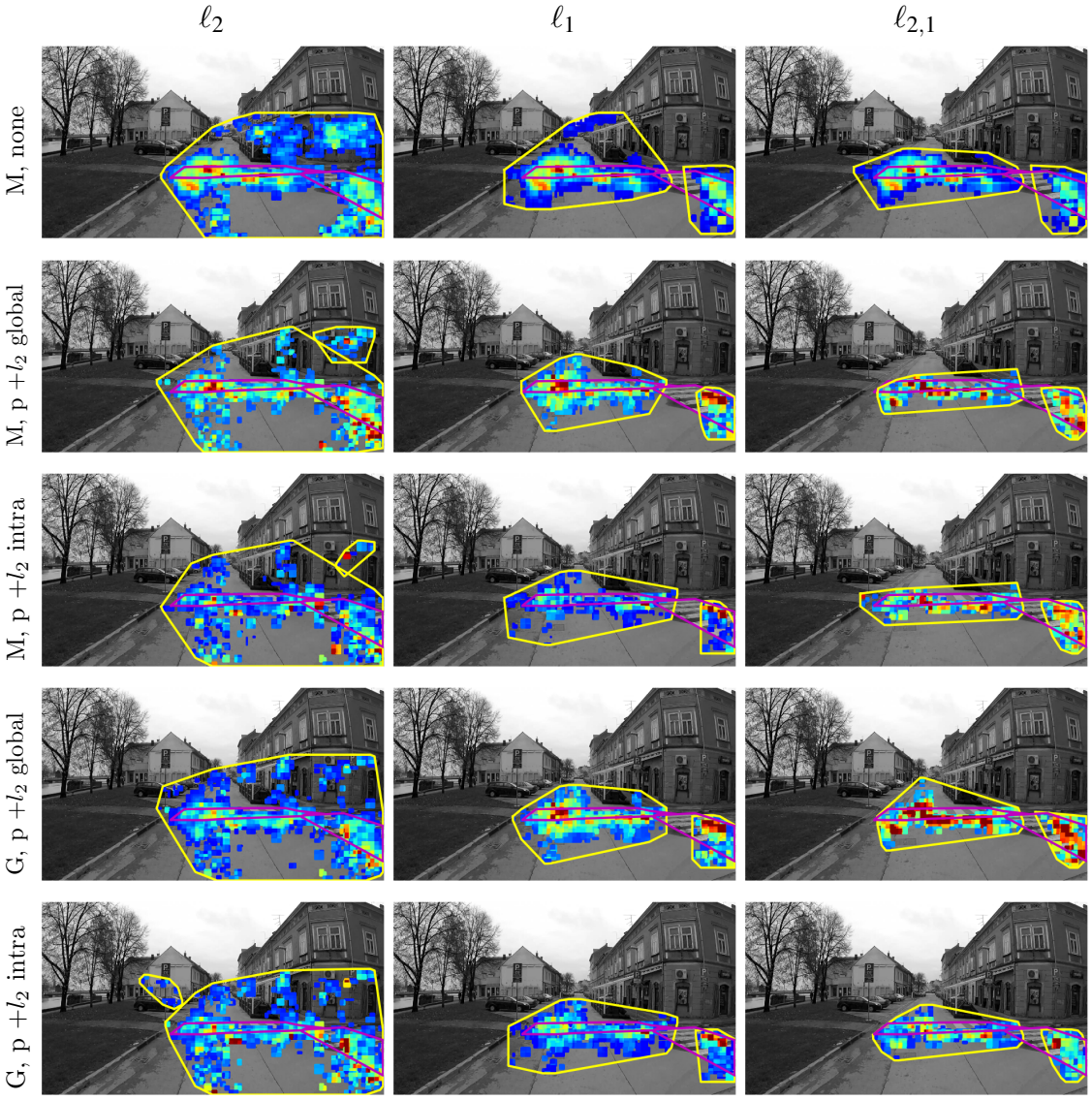
Br.	Konf.	FV Norm.	Regularizacija	*ACD*	*AP* testiranja	p_{miss}	t_{op}/s	
1	M (3.10)	-	ℓ_2	100	76	0.46	26.6	
2	M (3.10)	-	ℓ_1	78.9	91	0.33	19.1	
3	M (3.10)	-	$\ell_{2,1}$	14.8	92	0.31	2.8	
4	M (3.12)	p, ℓ_2 global	ℓ_1	30.5	90	0.30	52.7	
5	M (3.12)	p, ℓ_2 global	$\ell_{2,1}$	7.0	92	0.27	29.8	
6	M (3.12)	p, ℓ_2 intra	ℓ_1	21.1	87	0.28	10.8	
7	M (3.12)	p, ℓ_2 intra	$\ell_{2,1}$	4.7	**93**	**0.25**	0.8	
8	G (3.14)	p, ℓ_2 global	ℓ_1	30.5	87	0.31	6.8	7.8×
9	G (3.14)	p, ℓ_2 global	$\ell_{2,1}$	7.0	90	0.28	1.0	28.8×
10	G (3.14)	p, ℓ_2 intra	ℓ_1	21.1	89	0.29	3.8	2.9×
11	G (3.14)	p, ℓ_2 intra	$\ell_{2,1}$	4.7	**92**	**0.25**	**0.3**	2.7×

metričke normalizacije (ℓ_2 global) te izravnim proračunom odziva (retci 4 – 5) prema 3.12 smanjuje se frekvencija promašaja p_{miss} za do 4 postotna boda (redak 5 naspram retka 3). U slučaju $\ell_{2,1}$ regularizacije (krajnje desno u drugom retku slike 5.8) dobivaju se lokalizacijski poligoni vrlo bliski ciljanima (označenim purpurno-crvenom bojom).

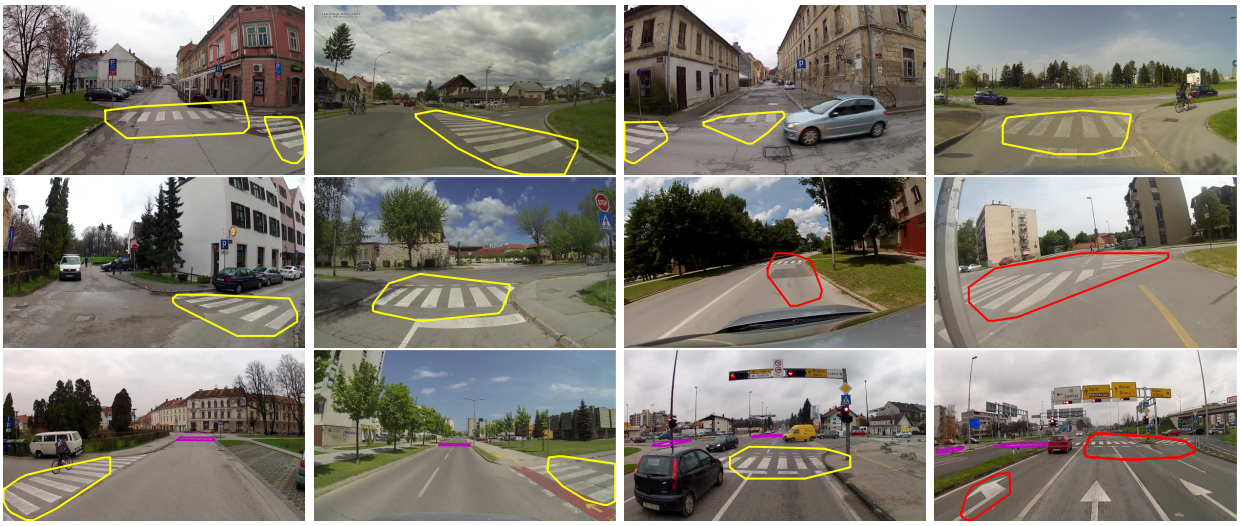
Rezultati dobiveni primjenom normalizacije potenciranjem (p) i metričke normalizacije po komponentama (ℓ_2 intra) ukazuju na činjenicu da ℓ_2 intra normalizacija u konjunkciji s $\ell_{2,1}$ rijetkim modelom daje najbolju lokalizacijsku točnost (93% *AP* testiranja i 0.25 p_{miss}, redak 7). S druge strane, rezultati također pokazuju da normalizacija po komponentama ima negativan utjecaj na lokalizacijsku učinkovitost kada se prilikom učenja primjenjuje ℓ_1 regularizacija koja ne uzima u obzir specifičnu blokovsku strukturu Fisherovih vektora (redak 6: pad od 4 posto *AP* u odnosu na redak 2).

Aproksimacija odziva okna primjenom gradijenta normaliziranog odziva slike (3.14) vrednuje se u sljedeće dvije skupine eksperimenata (retci 8 – 11). U slučaju globalne metričke normalizacije (retci 8 – 9), gradijent postiže tek nešto manju lokalizacijsku preciznost (pad do 3 posto u odnosu na retke 4 – 5). U slučaju metričke normalizacije po komponentama (redak 11), aproksimacija gradijentom daje sumjerljivu lokalizacijsku preciznost u slučaju $\ell_{2,1}$ modela. Peti redak slike 5.8 (krajnje desno) potvrđuje dobiveni rezultat. Primjenom ℓ_1 modela, slično kao

i u slučaju prometnih znakova, bilježi se poboljšanje lokalizacijske učinkovitosti u odnosu na izravan proračun (redak 10 naspram retka 6). Takvo je ponašanje posljedica činjenice da normalizacija po komponentama ima negativan utjecaj na modele koji ignoriraju strukturu Fisherovih vektora.



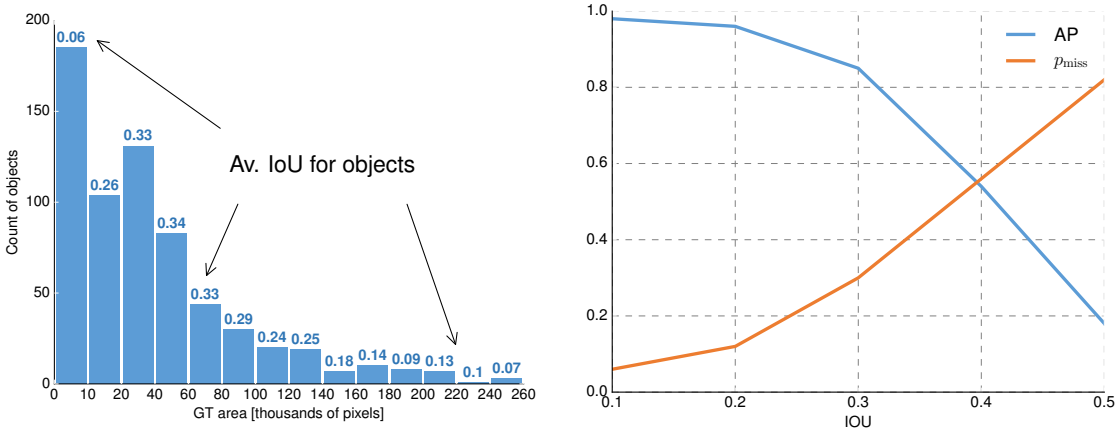
Slika 5.8: Usporedba ujedinjenih mapa odziva preko više mjerila za različite konfiguracije. Slika je prikazana u sivim tonovima kako bi se naglasile razlike u vrijednostima odziva, gdje odzivi rastu od tamno plave prema žutoj i naposlijetku crvenoj boji koja označava piksele najvećeg odziva. Slika prikazuje dva pješačka prijelaza, jedan u bočnoj poziciji u neposrednoj blizini kamere i drugi snimljen sprijeda udaljen desetak metara od kamere. Žutom bojom označeni su lokalizacijski poligoni dobiveni algoritmom 2 na temelju ujedinjenih mapa odziva za različite pristupe. Purpurno-crvenom (engl. *magenta*) bojom označeni su ciljani poligoni objekata označeni od strane ljudskog agenta (koriste se isključivo za vrednovanje učinkovitosti lokalizacije). Svaki redak odgovara skupini eksperimenata u tablici 5.6. Prvi stupac također pokazuje rezultate za ℓ_2 regularizirane modele koji nisu pogodni za rješavanje danog problema.



Slika 5.9: Rezultati lokalizacije pješačkih prijelaza na skupu za testiranje: valjani poligoni lokalizacije (zadovoljavaju uvjet (5.3)) prikazani su žutom bojom, a oni koji ne zadovoljavaju (5.3) prikazani su crveno. U slučajevima gdje lokalizacijski postupak ne uspijeva generirati poligon s nepraznim presjekom u odnosu na sam objekt, ručno označeni poligoni prikazani su purpurno-crvenom bojom.

5.3.5 Analiza neuspjelih lokalizacija pješačkih prijelaza

Slika 5.9 ilustrira primjere lokalizacijskih poligona dobivenih primjenom konfiguracije broj 7 u tablici 5.6. Poligoni koji zadovoljavaju uvjet (5.3) prikazani su žutom bojom. Rezultati pokazuju da predstavljena metoda uspješno lokalizira pješačke prijelaze snimljene iz različitih kutova gledišta (bočno i sprijeda) kao i izbijeljene pješačke prijelaze (redak 1, krajnje desno). Frekvencija promašaja za danu konfiguraciju iznosi $p_{\text{miss}} = 0.25$, a posljedica je dva osnovna tipa promašaja: i) promašaji lokalizacije izrazito malenih objekata udaljenih u odnosu na kameru te ii) promašaji lokalizacije međusobno bliskih objekata gdje algoritam generiranja lokalizacijskih poligona generira jedinstven poligon za bliske objekte. Analiza pokazuje da 80 posto promašaja odgovara malenim objektima udaljenim u odnosu na kameru (purpurno-crveni poligoni u trećem retku slike 5.9). Preostalih 20 posto promašaja uključuju poligone lokalizacije koji ne zadovoljavaju uvjet (5.3), a odnose se na bliske objekte koji se međusobno dodiruju (crveni poligoni u slikama krajnje desno u drugom i trećem retku slike 5.9). U tom slučaju, postupak za generiranje lokalizacijskih poligona na temelju ujedinjenih mapa odziva (algoritam 2) zbog međusobno bliskih piksela visokog odziva generira jedinstveni poligon za oba objekta, odnosno nije u stanju diferencirati između više objekata u slici. Prosječna površina promašenih objekata iznosi 4573 piksela, odnosno 0.5 posto ukupne površine slike. Kao što je već navedeno, takvi objekti nisu ključni za zadatak automatskog kartiranja cestovne infrastrukture jer se za njih ne može precizno odrediti GPS pozicija. Primjeri lažnih odziva (krajnje desno u trećem retku slike 5.9) uključuju elemente cestovne signalizacije koji sadrže slične uzorke kao i pješački prijelazi.



Slika 5.10: Utjecaj *IoU* praga (5.3) na učinkovitost lokalizacije pješačkih prijelaza za konfiguraciju koja odgovara devetom retku tablice 5.6. Lijevo: raspodjela objekata prema površini ručno označenog poligona. Za svaki odjeljak histograma prikazana je prosječna *IoU* vrijednost za objekte u tom intervalu. Desno: utjecaj *IoU* praga na prosječnu preciznost testiranja (*AP*, označeno plavom bojom) i frekvenciju promašaja (p_{miss}, označeno narančastom bojom) za objekte veće od 1 posto površine slike.

5.3.6 Analiza praga lokalizacije *IoU*

U okviru ovog odjeljka, razmatra se utjecaj hiper-parametra *IoU* koji odgovara omjeru presjeka i unije (engl. *Intersection over Union, IoU*) između procijenjenih i točnih lokalizacijskih poligona za konfiguraciju u devetom retku tablice 5.6 (*AP* = 93 posto, $p_{\text{miss}} = 0.25$). Slika 5.10 (lijevo) ilustrira raspodjelu veličine objekata u skupu za testiranje. Za svaki odjeljak histograma, prikazana je odgovarajuća prosječna *IoU* vrijednost dobivena vrednovanjem lokalizacijskih poligona za objekte u tom intervalu. Histogram ukazuje na činjenicu da je većina promašenih objekata (prosječna *IoU* vrijednost u iznosu 0.06) manja od 10^4 piksela, što znači da pokrivaju manje od 1 posto površine slike te su udaljeni u odnosu na kameru. Prosječne vrijednosti *IoU* praga za objekte veće od 10^4 piksela pokazuju da predstavljena metoda uspješno lokalizira barem određeni dio takvih objekata. Približno pola objekata (48 posto) lokalizirano je sa prosječnim *IoU* od 0.32. Navedeni objekti pokrivaju između 2 do 11 posto površine slike ($2 \cdot 10^4$ – 10^5 piksela). Za vrlo velike objekte ($\geq 2 \cdot 10^5$ piksela, gdje takvi objekti čine 2 posto ukupnog broja objekata u čitavom skupu podataka) bilježi se prosječna *IoU* vrijednost od 0.10. Budući da takvi objekti zauzimaju vrlo malen udio u skupu podataka, algoritam učenja ih ne uspijeva naučiti kao slikovni koncept.

Slika 5.10 (desno) prikazuje utjecaj *IoU* praga na prosječnu preciznost *AP* i frekvenciju promašaja p_{miss} za objekte veće od 10^4 piksela. Povećanjem *IoU* na vrijednost od 0.2, bilježi se manji pad vrijednosti *AP* mjere (u iznosu od 2 posto, sa 98 posto na 96 posto). Daljnjim povećanjem *IoU* na vrijednost 0.3, dobiva se preciznost u iznosu *AP* = 85 posto, dok frekvencija promašaja raste na vrijednost od 0.3. Rezultat dobiven uz *IoU* = 0.3 predstavlja razmjerno dobru lokalizacijsku učinkovitost s obzirom da se radi o modelu učenome uz slabi nadzor. Daljnje povećanje *IoU* praga značajno smanjuje preciznost lokalizacije *AP* zbog lažnih negativa i ne-

mogućnosti da algoritam generiranja lokalizacijskih poligona načini razliku između susjednih pješačkih prijelaza.

5.3.7 Analiza vremenskog izvođenja

U okviru ovog odjeljka detaljno se razmatra vremenska učinkovitost postupaka lokalizacije predstavljenih u tablici 5.6. Svi eksperimenti provedeni su na Intelovu procesoru E5-2620 frekvencije 2.00 GHz, pri čemu je korištena jedna jezgra.

Proces lokalizacije započinje izdvajanjem konvolucijskih značajki conv3-512 iz posljednjeg konvolucijskog sloja duboke VGG-E mreže. U prosjeku se iz svake slike izluči 6344 lokalna opisnika, a proces ekstrakcije značajki traje oko $t_{\mathrm{lf}} = 1\,\mathrm{s}$. Proračun vjerojatnosti pridruživanja u odnosu na komponente modela mješavine Gaussovih raspodjela traje u prosjeku $t_{\mathrm{sa}} = 0.11\,\mathrm{s}$. Dobivena vrijednost je 33 puta manja u odnosu na vrijeme mekog pridruživanja zabilježeno u slučaju prometnih znakova ($t_{\mathrm{sa}} = 3.7\,\mathrm{s}$) u okviru odjeljka 5.2.2. Zapaženo povećanje vremenske učinkovitosti izvođenja posljedica je nekoliko čimbenika:

- Broj izdvojenih lokalnih opisnika je 14 puta manji, nego u slučaju prometnih znakova (6344 naspram $87 \cdot 10^3$ slikovnih okana).
- Dimenzionalnost izdvojenih lokalnih opisnika je 6 puta veća u odnosu na prometne znakove (512 naspram 80).
- Broj komponenti slikovnog rječnika je 8 puta manji, nego za prometne znakove (128 naspram 1024 komponente).

Dobivene vrijednosti mjera t_{lf} i t_{sa} jednake su za sve lokalizacijske pristupe u tablici 5.6. Iz tog je razloga u nastavku posebna pažnja usmjerena na analizu vremena potrebnog za proračun odziva diskriminativnih okana t_{op}.

Ukupno vrijeme lokalizacije Prosječno vrijeme obrade slike ($t_{\mathrm{lf}} + t_{\mathrm{sa}} + t_{\mathrm{op}}$) u fazi lokalizacije iziskuje u prosjeku 1.9 s za lokalizacijski pristup prikazan u devetom retku tablice 5.6 (93 posto AP, 0.25 p_{miss}). Pri tome valja napomenuti da je sustav za lokalizaciju izveden u programskom jeziku Python [166]. Primjenom aproksimacije gradijentom za analogni skup normalizacija (redak 11 u tablici 5.6) vrijeme obrade dodatno se umanjuje na vrijednost od 1.4 s uz sumjerljivu lokalizacijsku preciznost. Dobiveno vrijeme izvođenja od 1.4 s vrlo je blisko izvođenju u stvarnom vremenu. Postignuti rezultat je značajan iz više razloga: i) navedena vremenska učinkovitost dobivena je modelom učenim u slabo nadziranom okruženju te ii) nisu primjenjivane metode segmentacije [44] ili objektnosti [134] kojima se zaobilazi pretraživanje čitavog prostora mogućih lokacija objekata. Obavljeno je globalno optimalno pretraživanje čitavog prostora mogućih hipoteza primjenom rijetkih modela i optimizacija proračuna odziva slikovnog okna opisanih u odjeljku 3.4.2.

Vrijeme proračuna odziva okna t_{op} Razmatra se usporedba vremenske učinkovitosti rijetkih modela u odnosu na guste ℓ_2 modele i aproksimacije gradijentom (3.14) u odnosu na izravan proračun odziva okna (3.12). Rezultati prikazani u tablici 5.6 ukazuju na to da rijetki $\ell_{2,1}$ modeli postižu ubrzanje od 10 puta u odnosu na guste ℓ_2 modele (redak 3 naspram retka 1), dok je ubrzanje u slučaju ℓ_1 modela manje (1.4 puta: redak 2 naspram retka 1). Prednost $\ell_{2,1}$ rijetkih modela u odnosu na ℓ_1 modele konzistentna je u svim eksperimentima, gdje se $\ell_{2,1}$ modelima postiže ubrzanje od 13 puta (redak 7 naspram retka 6, redak 11 naspram retka 10). Slično kao i u slučaju prometnih znakova, rezultati pokazuju da je globalna ℓ_2 normalizacija iznimno neefikasna u slučaju izravnog proračuna odziva okna (retci 4 – 5) jer zahtjeva proračun cjelokupnog Fisherova vektora. Izravan proračun odziva okna znatno je učinkovitiji u slučaju normalizacije po komponentama gdje se uslijed rijetkosti modela postiže ubrzanje u odnosu na slučaj bez normalizacija (retci 6, 7 naspram redaka 2, 3). Najbolje se performanse u vidu vremenske učinkovitosti postižu aproksimacijom gradijentom, koja postiže ubrzanje do 30 puta u slučaju globalne metričke normalizacije (retci 8, 9 naspram redaka 4, 5) i gotovo 3 puta u slučaju normalizacije po komponentama (retci 10, 11 naspram redaka 6, 7).

5.4 Rasprava eksperimentalnih rezultata

U posljednja dva poglavlja vrednovana je učinkovitost lokalizacije pristupom „odozdo prema gore" na dva različita skupa podataka. Usporedba rezultata predstavljenih u odjeljcima 5.2.2 i 5.3 ukazuje na četiri činjenice:

- Rijetki modeli istovremeno postižu bolju lokalizacijsku preciznost i vremensku učinkovitost izvođenja u odnosu na guste modele.
- Nelinearne normalizacije značajno pridonose učinkovitosti klasifikacije slika i lokalizacije objekata.
- Aproksimacija gradijentom postiže sumjerljivu lokalizacijsku točnost u odnosu na izravan proračun odziva okna te istovremeno značajno povećava vremensku učinkovitost izvođenja.
- Metrička normalizacija po komponentama umanjuje frekvenciju promašaja p_{miss} u slučaju kada se model uči uz primjenu $\ell_{2,1}$ regularizacije.

U slučaju kada se koristi ℓ_1 rijetka regularizacijska funkcija, koja ne uzima u obzir blokovsku strukturu Fisherovih vektora, metrička normalizacija po komponentama ima negativan utjecaj na lokalizacijsku preciznost.

Eksperimenti također pokazuju značajne razlike između dva skupa podataka. Za zadatak klasifikacije slika, rijetki modeli postižu bolju klasifikacijsku preciznost za prometne znakove (do 9 postotnih bodova), dok se u slučaju pješačkih prijelaza bilježi podjednaka klasifikacijska učinkovitost za rijetke i guste modele. Za zadatak lokalizacije, rijetki modeli postižu bolju

preciznost na oba skupa podataka. U slučaju pješačkih prijelaza, slike uz pješačke prijelaze obično sadrže i objekte poput prometnih znakova, svjetlosnih znakova (semafora) te ostale elemente cestovne signalizacije. Gusti modeli, stoga, postižu podjednaku klasifikacijsku točnost fokusirajući se na kontekst, a ne nužno na objekte traženog razreda.

Klasifikacijska učinkovitost za pješačke prijelaze bolja je nego u slučaju prometnih znakova (98 posto naspram 81 posto). Dobiveni rezultat posljedica je sljedećeg:

- Pješački su prijelazi veći u odnosu na prometne znakove te kao takvi daju veći doprinos cjelokupnom opisniku slike.
- Drugi objekti koji se pojavljuju uz pješačke prijelaze u pozitivnim slikama doprinose povećanju klasifikacijske točnosti.

Objekti koji se najčešće pojavljuju uz pješačke prijelaze otežavaju zadatak slabo nadzirane lokalizacije. Rezultati prikazani u tablici 5.6 i slici 5.8 ukazuju na činjenicu da rijetki modeli uspješno identificiraju slikovna okna na pješačkim prijelazima, dok gusti modeli nisu u stanju ignorirati ostale objekte karakteristične za pozitivne slike.

Eksperimenti s nelinearnim normalizacijama pokazuju da normalizacije poboljšavaju preciznost lokalizacije AP i umanjuju frekvenciju promašaja p_{miss} u slučaju prometnih znakova, dok u slučaju pješačkih prijelaza umanjuju frekvenciju promašaja i postižu sumjerljivu lokalizacijsku preciznost.

Ukupno vrijeme izvođenja 3 puta je sporije u slučaju prometnih znakova. Temeljni uzrok ove značajne razlike leži u vremenu potrebnom za proračun vjerojatnosti pridruživanja u odnosu na komponente slikovnog rječnika (2.23) koje je 33 puta sporije u slučaju prometnih znakova. Jedno moguće rješenje opisanog problema leži u primjeni metode ubrzanog mekog pridruživanja predložene u okviru rada [167].

U slučaju lokalizacije primjenom opisnika prostornog rasporeda (odjeljak 5.2.3) postiže se bolja klasifikacijska i lokalizacijska preciznost u odnosu na modele temeljene na značajkama izgleda. Usporedba prostornih histograma i Fisherovih vektora pokazuje oba modela postižu sumjerljivu klasifikacijsku preciznost. Za zadatak lokalizacije, prostorni Fisherovi vektori postižu značajno bolju lokalizacijsku točnost (porast AP za šest postotnih bodova uz istovremeno smanjenje p_{miss} za tri postotna boda). Razlog tome leži u činjenici da prostorni histogram ne uzima u obzir raspodjelu unutar pojedinog odjeljka. Eksperimenti pokazuju da prostorni Fisherovi vektori ne povećavaju složenost izvođenja budući da se vrijednosti Fisherovih vektora mogu unaprijed izračunati i pohraniti u priručnu tablicu.

Poglavlje 6

Zaključak

Detekcija prisutnosti i lokacije objekata u slikama važni su zadaci u računalnog vida. Tijekom posljednjeg desetljeća zapažen je značajan napredak na području lokalizacije objekata i klasifikacije slika, a najbolji su rezultati postignuti postupcima strogo nadziranog učenja. Učenje lokalizacijskih modela postupcima strogo nadziranog učenja zahtijeva oznake lokacija objekata u vidu opisanih pravokutnika ili poligona na razini točnosti pojedinih piksela. S druge strane, svakodnevno se na Internet poslužitelje pohranjuju velike količine slika i video zapisa. Goleme količine slikovnih sadržaja dostupne na Internetu karakterizirane su velikim brojem objekata različitih razreda, a opisane su tekstualnim oznakama na razini slike. Primjer takvih oznaka uključuje: „#automobil, #semafor, #autobus", gdje potonji niz imenica označava da se u slici nalaze objekti razreda automobil, semafor i autobus. U općenitom slučaju, za takve slike ne postoji informacija o položaju pojedinih objekata. S obzirom na velike količine podataka i razreda gotovo je nemoguće označiti sve objekte. Iz navedenih razloga, javlja se potreba za razvojem metoda strojnog učenja i računalnog vida koje bi omogućile razumijevanje tako označenog slikovnog sadržaja. Jedan od mogućih pristupa problemu nedostajućih oznaka objekata jest u vidu paradigma slabo nadziranog učenja.

U okviru ovog rada predstavljen je pristup slabo nadziranoj lokalizaciji objekata temeljen na načelu „odozdo prema gore": od primitivnih slikovnih okana ka složenijim slikovnim strukturama. Glavni ciljevi prilikom dizajna sustava lokalizacije bili su: i) razviti sustav koji uspješno lokalizira objekte proizvoljnih veličina u složenim scenama s obzirom na trend ubrzanog rasta slabo označenih slikovnih podataka te ii) razviti sustav koji navedeni zadatak lokalizacije učinkovito obavlja sa stajališta vremenske učinkovitosti. S obzirom na potencijalno velik broj mogućih lokacija objekata, navedeni ciljevi međusobno su komplementarni, a zadatak realizacije zanimljiv i težak.

Problem neoznačenih lokacija objekata u podacima za učenje rješava se učenjem rijetkih lokalizacijskih modela na temelju Fisherovih vektora cjelokupnih slika i oznaka prisutnosti objekata u slikama. Aditivnost reprezentacije Fisherovih vektora i svojstvo da Fisherovi vek-

tori poništavaju utjecaj pozadinske informacije u slici omogućuju učenje u slabo nadziranom okruženju, odnosno ostvarenje prvog cilja zadanog prilikom dizajniranja sustava. Za ostvarenje drugog cilja koriste se modeli rijetki po komponentama. Navedeni se modeli dobivaju učenjem uz primjenu $\ell_{2,1}$ regularizacijske funkcije čime se smanjuje i) utjecaj prenaučenosti te ii) dimezionalnost modela identifikacijom diskriminativnih slikovnih riječi. Primjena modela rijetkih po komponentama značajno poboljšava preciznost lokalizacije u odnosu na guste ℓ_2 modele (18 postotnih bodova za prometne znakove i 16 postotnih bodova za pješačke prijelaze). Istovremeno se postiže i značajno ubrzanje budući da po komponentama rijetki modeli koriste samo djelić cjelokupne visokodimenzionalne reprezentacije (1 posto za prometne znakove i oko 15 posto za pješačke prijelaze).

U svrhu poboljšanja lokalizacijske točnosti, u sustav lokalizacije „odozdo prema gore" uvode se nelinearne normalizacije Fisherovih vektora cjelokupnih slika. Glavni izazov prilikom izvedbe navedenog zadataka bio je postići vremensku učinkovitost proračuna odziva slikovnog okna. Normalizacije invalidiraju aditivnost slikovnih okana i sprečavaju primjenu modela učenog nad normaliziranim slikama na razini slikovnog okna. Iz navedenog je razloga predložena aproksimacija odziva skalarnim produktom gradijenta odziva normalizirane slike i Fisherova vektora okna. Eksperimenti pokazuju da predložena aproksimacija rezultira sumjerljivom lokalizacijskom točnošću te značajnim ubrzanjem (oko 200 puta na skupu podataka za prometne znakove). Za zadatak lokalizacije pješačkih prijelaza ustanovljena je vremenska učinkovitost bliska izvođenju u stvarnom vremenu od 1.9 sekundi. Dobiveni rezultat je značajan budući da se radi o primjeni modela učenog u paradigmi slabog nadzora. Pokazuje se da se učenjem lokalizacijskih modela isključivo na temelju oznaka prisutnosti objekata u slikama može postići i lokalizacijska točnost i vremenska učinkovitost. Uz aproskimaciju odziva okana pomoću gradijenta, u opisani sustav uvodi se i metrička normalizacija po komponentama. Uporaba metričke normalizacije po komponentama u konjunkciji sa po komponentama rijetkim modelima dodatno povećava stupanj rijetkosti modela i smanjuje frekvenciju promašaja.

S obzirom da reprezentacija Fisherovim vektorima ne uzima u obzir prostorne odnose slikovnih okana, razmatrane su slikovne reprezentacije prostornog rasporeda okana. Razvijena su dva opisnika za predstavljanje lokalnog prostornog rasporeda kojima se modeliraju prostorne konstelacije slikovnih riječi: prostorni histogrami i prostorni Fisherovi vektori. Eksperimenti na području lokalizacije prometnih znakova pokazuju da se primjenom opisnika prostornog rasporeda uklanjaju mnogi lažni odzivi i shodno tome povećava lokalizacijska točnost (porast preciznosti za 4 postotna boda i pad frekvencije promašaja za čak 5 postotnih bodova za prostorne Fisherove vektore). Dodatna prednost reprezentacija prostornog rasporeda jest da se lokalizacijski poligoni formiraju na temelju svih okana pozitivnog odziva, čime se smanjuje broj parametara prilikom generiranja lokalizacijskih poligona.

Iscrpna eksperimentalna evaluacija predloženog sustava lokalizacije provedena je za zadatke

klasifikacije slika i lokalizacije objekata. Pokazano je da se metode slabo nadziranog učenja mogu primijeniti u svrhu automatizacije provjera cestovne infrastrukture i automatizacije digitalnog kartiranja. U okviru zadatka automatizacije kartiranja prikupljen je nov skup podataka od 2381 slike pješačkih prijelaza. Navedeni skup podataka dobiven je poluautomatskim uparivanjem lokacija objekata iz OpenStreetMap karte u odnosu na georeferencirani video. Za potrebe automatizacije cestovnih inspekcija korišten je javno dostupan skup podataka za prometne znakove. Rezultati pokazuju da se predloženim pristupom uspješno mogu lokalizirati izrazito maleni objekti poput prometnih znakova, ali i veliki objekti sa velikom unutar razrednom varijacijom poput pješačkih prijelaza. Brojčani pokazatelji lokalizacije bolji su u slučaju prometnih znakova, gdje se uz IoU od 50 posto i SIFT značajke postiže preciznost od 86 posto i frekvencija promašaja od 0.12. U slučaju pješačkih prijelaza, uz IoU u iznosu od 10 posto i konvolucijske značajke postignuta je preciznost od 92 posto i frekvencija promašaja od 0.25. Analizom neuspjelih slučajeva lokalizacije pješačkih prijelaza ustanovljeno je da su najčešći uzroci promašaja i) izrazito maleni objekti udaljeni u odnosu na kameru te ii) višestruki bliski objekti za koje algoritam generiranja lokalizacijskih poligona generira jedan poligon. Ipak, glavni cilj automatizacije digitalne kartografije jest lokalizirati objekte relativno bliske u odnosu na kameru. Za takve se objekte, uz pomoć prostornog-vremenskih podataka vezanih uz georeferencirani video, mogu s razmjernom pouzdanošću odrediti GPS koordinate. Budući da se za malene objekte udaljene u odnosu na kameru ne mogu odrediti GPS koordinate s odgovarajućom pouzdanosti, ocijenjeno je da njihova lokalizacija nije ključna za automatizaciju digitalne kartografije.

U nastavku su opisana moguća poboljšanja postojećeg sustava i pravci daljnjeg istraživanja. Zanimljiv smjer budućeg rada uključuje mogućnost primjene nelinearnih normalizacija za modeliranje lokalnog prostornog rasporeda slike. U okviru drugog doprinosa ove disertacije pokazano je da normalizacije povećavaju lokalizacijsku točnost modela prve razine. Iz tog se razloga i na području prostornog rasporeda mogu očekivati slična poboljšanja. Nadalje, pokazano je da je predstavljeni pristup kompatibilan sa različitim tipovima značajki niske razine. Za problem lokalizacije prometnih znakova korištene su SIFT značajke, dok su prilikom lokalizacije pješačkih prijelaza korištene konvolucijske značajke. Kako bi se ustanovilo koje su značajke najprikladnije za rješavanje navedenih problema, razmatrat će se proširenje eksperimentalnog vrednovanja korištenjem oba tipa značajki (SIFT za pješačke prijelaze, konvolucijske značajke za prometne znakove). Posebno zanimljiv smjer istraživanja uključuje potpunu automatizaciju procesa digitalne kartografije. U okviru ove disertacije predloženo je poluautomatsko prikupljanje slabo označenog skupa podataka na temelju georeferenciranog videa i OpenStreetMap lokacija. Postupkom potpuno automatskog mapiranja OSM lokacija i prostorno-vremenskih podataka vezanih uz video, uz slike objekata dobiveno je i 15 posto slika koje ne sadrže objekte od interesa. Dobivene slike uklonjene su ručnim probirom iz skupa za učenje. Poseban izazov

zadatka lokalizacije bio bi, stoga, zaobići postupak ručnog probira dobivenih slika i naučiti nove slikovne koncepte na temelju zašumljenih oznaka. Naposljetku, s obzirom da učenje „s kraja na kraj" postiže iznimno dobre rezultate na različitim područjima računalnog vida, razmatrat će se njegova primjena za zadatak slabo nadzirane lokalizacije.

Literatura

[1] Verbeek, J., Schmid, C., "Predavanja iz predmeta strojno učenje i raspoznavanje objekata", dostupno na: http://lear.inrialpes.fr/people/verbeek/MLOR.16.17.php (14. prosinca 2016.).

[2] Zhang, N., Donahue, J., Girshick, R. B., Darrell, T., "Part-based R-CNNs for fine-grained category detection", in Computer Vision - ECCV 2014 - 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part I, 2014, str. 834–849, dostupno na: http://dx.doi.org/10.1007/978-3-319-10590-1_54

[3] Viola, P. A., Jones, M. J., "Robust real-time face detection", International Journal of Computer Vision, Vol. 57, No. 2, 2004, str. 137–154, dostupno na: http://dx.doi.org/10.1023/B:VISI.0000013087.49260.fb

[4] Dalal, N., Triggs, B., "Histograms of oriented gradients for human detection", in CVPR, 2005.

[5] Felzenszwalb, P. F., Girshick, R. B., McAllester, D. A., Ramanan, D., "Object detection with discriminatively trained part-based models", IEEE Trans. Pattern Anal. Mach. Intell., Vol. 32, No. 9, 2010, str. 1627–1645, dostupno na: http://dx.doi.org/10.1109/TPAMI.2009.167

[6] Cinbis, R. G., Verbeek, J. J., Schmid, C., "Segmentation driven object detection with Fisher Vectors", in IEEE International Conference on Computer Vision, ICCV 2013, Sydney, Australia, December 1-8, 2013, 2013, str. 2968–2975, dostupno na: http://dx.doi.org/10.1109/ICCV.2013.369

[7] Krizhevsky, A., Sutskever, I., Hinton, G. E., "ImageNet classification with deep convolutional neural networks", in Advances in Neural Information Processing Systems 25: 26th Annual Conference on Neural Information Processing Systems 2012. Proceedings of a meeting held December 3-6, 2012, Lake Tahoe, Nevada, United States., 2012, str. 1106–1114, dostupno na: http://papers.nips.cc/paper/4824-imagenet-classification-with-deep-convolutional-neural-networks

[8] Russakovsky, O., “Scaling up object detection”, Doktorski rad, Stanford University, Stanford, CA, SAD, 2015.

[9] Denton, E., Weston, J., Paluri, M., Bourdev, L. D., Fergus, R., “User conditional hashtag prediction for images”, in Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Sydney, NSW, Australia, August 10-13, 2015, 2015, str. 1731–1740, dostupno na: http://doi.acm.org/10.1145/2783258.2788576

[10] Zhang, W., Zeng, S., Wang, D., Xue, X., “Weakly supervised semantic segmentation for social images”, in IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015, 2015, str. 2718–2726, dostupno na: http://dx.doi.org/10.1109/CVPR.2015.7298888

[11] Cho, M., Kwak, S., Schmid, C., Ponce, J., “Unsupervised object discovery and localization in the wild: Part-based matching with bottom-up region proposals”, in IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015, 2015, str. 1201–1210, dostupno na: http://dx.doi.org/10.1109/CVPR.2015.7298724

[12] Doersch, C., Gupta, A., Efros, A. A., “Unsupervised visual representation learning by context prediction”, in 2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015, 2015, str. 1422–1430, dostupno na: http://dx.doi.org/10.1109/ICCV.2015.167

[13] Tang, Y., Wang, J., Gao, B., Dellandréa, E., Gaizauskas, R. J., Chen, L., “Large scale semi-supervised object detection using visual and semantic knowledge transfer”, in 2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016, 2016, str. 2119–2128, dostupno na: http://doi.ieeecomputersociety.org/10.1109/CVPR.2016.233

[14] Chapelle, O., Schölkopf, B., Zien, A., Semi-supervised learning, ser. Adaptive computation and machine learning. Cambridge (Mass.): MIT Press, 2006.

[15] Käding, C., Freytag, A., Rodner, E., Bodesheim, P., Denzler, J., “Active learning and discovery of object categories in the presence of unnameable instances”, in IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015, 2015, str. 4343–4352, dostupno na: http://dx.doi.org/10.1109/CVPR.2015.7299063

[16] Vijayanarasimhan, S., Grauman, K., “Large-scale live active learning: Training object detectors with crawled data and crowds”, International Journal of Computer

Vision, Vol. 108, No. 1-2, 2014, str. 97–114, dostupno na: http://dx.doi.org/10.1007/s11263-014-0721-9

[17] Elahi, M., Ricci, F., Rubens, N., "A survey of active learning in collaborative filtering recommender systems", Computer Science Review, Vol. 20, 2016, str. 29–50, dostupno na: http://dx.doi.org/10.1016/j.cosrev.2016.05.002

[18] Cinbis, R. G., Verbeek, J. J., Schmid, C., "Weakly supervised object localization with multi-fold multiple instance learning", CoRR, Vol. abs/1503.00949, 2015, dostupno na: http://arxiv.org/abs/1503.00949

[19] Hoffman, J., Pathak, D., Darrell, T., Saenko, K., "Detector discovery in the wild: Joint multiple instance and representation learning", in IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015, 2015, str. 2883–2891, dostupno na: http://dx.doi.org/10.1109/CVPR.2015.7298906

[20] Zadrija, V., Krapac, J., Verbeek, J. J., Šegvić, S., "Patch-level spatial layout for classification and weakly supervised localization", in Pattern Recognition - 37th German Conference, GCPR 2015, Aachen, Germany, October 7-10, 2015, Proceedings, 2015, str. 492–503, dostupno na: http://dx.doi.org/10.1007/978-3-319-24947-6_41

[21] Wang, C., Ren, W., Huang, K., Tan, T., "Weakly supervised object localization with latent category learning", in Computer Vision - ECCV 2014 - 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part VI, 2014, str. 431–445, dostupno na: http://dx.doi.org/10.1007/978-3-319-10599-4_28

[22] Crowley, E., Zisserman, A., "Of gods and goats: Weakly supervised learning of figurative art", in British Machine Vision Conference, BMVC 2013, Bristol, UK, September 9-13, 2013, 2013, dostupno na: http://dx.doi.org/10.5244/C.27.39

[23] Chum, O., Zisserman, A., "An exemplar model for learning object classes", in 2007 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2007), 18-23 June 2007, Minneapolis, Minnesota, USA, 2007, dostupno na: http://dx.doi.org/10.1109/CVPR.2007.383050

[24] Galleguillos, C., Babenko, B., Rabinovich, A., Belongie, S. J., "Weakly supervised object localization with stable segmentations", in Computer Vision - ECCV 2008, 10th European Conference on Computer Vision, Marseille, France, October 12-18, 2008, Proceedings, Part I, 2008, str. 193–207, dostupno na: http://dx.doi.org/10.1007/978-3-540-88682-2_16

[25] Siva, P., Xiang, T., “Weakly supervised object detector learning with model drift detection”, in IEEE International Conference on Computer Vision, ICCV 2011, Barcelona, Spain, November 6-13, 2011, 2011, str. 343–350, dostupno na: http://dx.doi.org/10.1109/ICCV.2011.6126261

[26] Deselaers, T., Alexe, B., Ferrari, V., “Weakly supervised localization and learning with generic knowledge”, International Journal of Computer Vision, Vol. 100, No. 3, 2012, str. 275–293, dostupno na: http://dx.doi.org/10.1007/s11263-012-0538-3

[27] Bilen, H., Pedersoli, M., Tuytelaars, T., “Weakly supervised object detection with convex clustering”, in IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015, 2015, str. 1081–1089, dostupno na: http://dx.doi.org/10.1109/CVPR.2015.7298711

[28] Bilen, H., Vedaldi, A., “Weakly supervised deep detection networks”, in 2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016, 2016, str. 2846–2854, dostupno na: http://dx.doi.org/10.1109/CVPR.2016.311

[29] Diba, A., Sharma, V., Pazandeh, A. M., Pirsiavash, H., Gool, L. V., “Weakly supervised cascaded convolutional networks”, CoRR, Vol. abs/1611.08258, 2016, dostupno na: http://arxiv.org/abs/1611.08258

[30] Nguyen, M. H., Torresani, L., la Torre, F. D., Rother, C., “Learning discriminative localization from weakly labeled data”, Pattern Recognition, Vol. 47, No. 3, 2014, str. 1523–1534, dostupno na: http://dx.doi.org/10.1016/j.patcog.2013.09.028

[31] Perronnin, F., Dance, C. R., “Fisher kernels on visual vocabularies for image categorization”, in 2007 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2007), 18-23 June 2007, Minneapolis, Minnesota, USA, 2007, dostupno na: http://dx.doi.org/10.1109/CVPR.2007.383266

[32] Perronnin, F., Sánchez, J., Mensink, T., “Improving the Fisher kernel for large-scale image classification”, in ECCV, 2010, str. 143-156.

[33] Sánchez, J., Perronnin, F., Mensink, T., Verbeek, J. J., “Image classification with the Fisher vector: Theory and practice”, International Journal of Computer Vision, Vol. 105, No. 3, 2013, str. 222–245, dostupno na: http://dx.doi.org/10.1007/s11263-013-0636-x

[34] Murphy, K., Machine learning : a probabilistic perspective. Cambridge, Mass: MIT Press, 2012.

[35] Jenatton, R., Mairal, J., Obozinski, G., Bach, F. R., "Proximal methods for hierarchical sparse coding", Journal of Machine Learning Research, Vol. 12, 2011, str. 2297–2334.

[36] Yuan, M., Lin, Y., "Model selection and estimation in regression with grouped variables", Journal of the Royal Statistical Society: Series B (Statistical Methodology), Vol. 68, No. 1, 2006, str. 49–67.

[37] Lowe, D. G., "Distinctive image features from scale-invariant keypoints", International Journal of Computer Vision, Vol. 60, No. 2, 2004, str. 91–110, dostupno na: http://dx.doi.org/10.1023/B:VISI.0000029664.99615.94

[38] Cimpoi, M., Maji, S., Kokkinos, I., Vedaldi, A., "Deep filter banks for texture recognition, description, and segmentation", International Journal of Computer Vision, Vol. 118, No. 1, 2016, str. 65–94, dostupno na: http://dx.doi.org/10.1007/s11263-015-0872-3

[39] Simonyan, K., Zisserman, A., "Very deep convolutional networks for large-scale image recognition", CoRR, Vol. abs/1409.1556, 2014, dostupno na: http://arxiv.org/abs/1409.1556

[40] Csurka, G., Dance, C. R., Fan, L., Willamowski, J., Bray, C., "Visual categorization with bags of keypoints", in In Workshop on Statistical Learning in Computer Vision, ECCV, 2004, str. 1–22.

[41] Sivic, J., Zisserman, A., "Video Google: A text retrieval approach to object matching in videos", in 9th IEEE International Conference on Computer Vision (ICCV 2003), 14-17 October 2003, Nice, France, 2003, str. 1470–1477, dostupno na: http://dx.doi.org/10.1109/ICCV.2003.1238663

[42] Bishop, C., Pattern recognition and machine learning. New York: Springer, 2006.

[43] Jaakkola, T. S., Haussler, D., "Exploiting generative models in discriminative classifiers", in Advances in Neural Information Processing Systems 11, [NIPS Conference, Denver, Colorado, USA, November 30 - December 5, 1998], 1998, str. 487–493, dostupno na: http://papers.nips.cc/paper/1520-exploiting-generative-models-in-discriminative-classifiers

[44] Uijlings, J. R. R., van de Sande, K. E. A., Gevers, T., Smeulders, A. W. M., "Selective search for object recognition", International Journal of Computer Vision, Vol. 104, No. 2, 2013, str. 154–171, dostupno na: http://dx.doi.org/10.1007/s11263-013-0620-5

[45] Arandjelovic, R., Zisserman, A., "All about VLAD", in 2013 IEEE Conference on Computer Vision and Pattern Recognition, Portland, OR, USA, June 23-28, 2013, 2013, str. 1578–1585, dostupno na: http://dx.doi.org/10.1109/CVPR.2013.207

[46] Jegou, H., Douze, M., Schmid, C., "On the burstiness of visual elements", in 2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2009), 20-25 June 2009, Miami, Florida, USA, 2009, str. 1169–1176, dostupno na: http://dx.doi.org/10.1109/CVPRW.2009.5206609

[47] Oliva, A., Torralba, A., "Modeling the shape of the scene: A holistic representation of the spatial envelope", International Journal of Computer Vision, Vol. 42, No. 3, 2001, str. 145–175, dostupno na: http://dx.doi.org/10.1023/A:1011139631724

[48] Khoreva, A., Benenson, R., Omran, M., Hein, M., Schiele, B., "Weakly supervised object boundaries", in 2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016, 2016, str. 183–192, dostupno na: http://dx.doi.org/10.1109/CVPR.2016.27

[49] Li, D., Huang, J., Li, Y., Wang, S., Yang, M., "Weakly supervised object localization with progressive domain adaptation", in 2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016, 2016, str. 3512–3520, dostupno na: http://dx.doi.org/10.1109/CVPR.2016.382

[50] Oquab, M., Bottou, L., Laptev, I., Sivic, J., "Is object localization for free? - weakly-supervised learning with convolutional neural networks", in IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015, 2015, str. 685–694, dostupno na: http://dx.doi.org/10.1109/CVPR.2015.7298668

[51] Elvik, R., The handbook of road safety measures. Bingley, UK: Emerald, 2009.

[52] Kumar, P., Cahalane, C., McCarthy, T., "Initial results from European Road Safety Inspection (EURSI) mobile mapping project", in ISPRS Commission V Technical Symposium, Vol. 2124, Newcastle, UK, 2010.

[53] Šegvić, S., Brkić, K., Kalafatić, Z., Stanisavljević, V., Ševrović, M., Budimir, D., Dadić, I., "A computer vision assisted geoinformation inventory for traffic infrastructure", in 13th International IEEE Conference on Intelligent Transportation Systems, Sept 2010, str. 66-73.

[54] Gonzalez, Á., Garrido, M. Á., Llorca, D. F., Gavilán, M., Fernandez, J. P., Alcantarilla, P. F., Parra, I., Herranz, F., Bergasa, L. M., Sotelo, M. Á. G., de Toro, P. R., "Automatic traffic signs and panels inspection system using computer vision", IEEE

Trans. Intelligent Transportation Systems, Vol. 12, No. 2, 2011, str. 485–499, dostupno na: http://dx.doi.org/10.1109/TITS.2010.2098029

[55] Kumar, P., McElhinney, C. P., Lewis, P., McCarthy, T., "Automated road markings extraction from mobile laser scanning data", Int. J. Applied Earth Observation and Geoinformation, Vol. 32, 2014, str. 125–137, dostupno na: http://dx.doi.org/10.1016/j.jag.2014.03.023

[56] Maldonado-Bascon, S., Lafuente-Arroyo, S., Siegmann, P., Gomez-Moreno, H., Acevedo-Rodriguez, F. J., "Traffic sign recognition system for inventory purposes", in 2008 IEEE Intelligent Vehicles Symposium, June 2008, str. 590-595.

[57] Hou, Z., An Automated Road Sign Inventory System Based on Computer Vision. University of Arkansas, 2009.

[58] Frančula, N., Digitalna kartografija, Geodetski fakultet, Sveučilište u Zagrebu, Zagreb, 2004.

[59] Hake, G., Kartographie. Berlin, New York: De Gruyter, 1994.

[60] Google Inc, "Google Maps", dostupno na: https://maps.google.com/ (05. prosinca 2016.).

[61] Google Inc, "Google Earth", dostupno na: https://earth.google.com/ (05. prosinca 2016.).

[62] OpenStreetMap Foundation, "OpenStreetMap", dostupno na: https://www.openstreetmap.org (14. studenog 2016.).

[63] OpenStreetMap Wiki, "Bing maps — OpenStreetMap Wiki,", dostupno na: http://wiki.openstreetmap.org/w/index.php?title=Bing_Maps&oldid=1339673 (05. prosinca 2016.).

[64] OpenStreetMap Wiki, "Stats — OpenStreetMap Wiki,", dostupno na: http://wiki.openstreetmap.org/w/index.php?title=Stats&oldid=1394217 (5. prosinca 2016.).

[65] OpenStreetMap Wiki, "Recording GPS tracks — OpenStreetMap Wiki,", dostupno na: http://wiki.openstreetmap.org/w/index.php?title=Recording_GPS_tracks&oldid=1252513 (5. prosinca 2016.).

[66] Telenav GmbH, "OpenStreetView mobile application", dostupno na: https://play.google.com/store/apps/details?id=com.telenav.streetview&hl=en (16. studenog 2016.).

[67] Mapillary, “Mapillary mobile application”, dostupno na: https://play.google.com/store/apps/details?id=app.mapillary&hl=en (11. studenog 2016.).

[68] Wikipedia, “Smart city — wikipedia, the free encyclopedia”, dostupno na: https://en.wikipedia.org/w/index.php?title=Smart_city&oldid=751401069 (25. studenoga 2016.).

[69] RTA-ITS, “RTA Smart Drive mobile application”, dostupno na: https://play.google.com/store/apps/details?id=com.mireo.rtasmartdrive&hl=en (05. prosinca 2016.).

[70] Promet i Prostor, “E-roads web platform”, dostupno na: https://prometiprostor.hr/en/solutions/ (19. rujna 2016.).

[71] Andrews, S., Tsochantaridis, I., Hofmann, T., “Support vector machines for multiple-instance learning”, in NIPS, 2003.

[72] Šegvić, S., Brkić, K., Kalafatić, Z., Pinz, A., “Exploiting temporal and spatial constraints in traffic sign detection from a moving vehicle”, Mach. Vis. Appl., Vol. 25, No. 3, 2014, str. 649–665, dostupno na: http://dx.doi.org/10.1007/s00138-011-0396-y

[73] Krapac, J., Šegvić, S., “Predavanja iz predmeta “Modeli za predstavljanje slike i videa””, dostupno na: http://www.fer.unizg.hr/predmet/mzpsv (8. prosinca 2016.).

[74] Jegou, H., Douze, M., Schmid, C., Pérez, P., “Aggregating local descriptors into a compact image representation”, in The Twenty-Third IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2010, San Francisco, CA, USA, 13-18 June 2010, 2010, str. 3304–3311, dostupno na: http://dx.doi.org/10.1109/CVPR.2010.5540039

[75] Lazebnik, S., Schmid, C., Ponce, J., “Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories”, in 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2006), 17-22 June 2006, New York, NY, USA, 2006, str. 2169–2178, dostupno na: http://dx.doi.org/10.1109/CVPR.2006.68

[76] van Gemert, J. C., Veenman, C. J., Smeulders, A. W. M., Geusebroek, J., “Visual word ambiguity”, IEEE Trans. Pattern Anal. Mach. Intell., Vol. 32, No. 7, 2010, str. 1271–1283, dostupno na: http://dx.doi.org/10.1109/TPAMI.2009.132

[77] Nowak, E., Jurie, F., Triggs, B., “Sampling strategies for bag-of-features image classification”, in Computer Vision - ECCV 2006, 9th European Conference on Computer Vision, Graz, Austria, May 7-13, 2006, Proceedings, Part IV, 2006, str. 490–503, dostupno na: http://dx.doi.org/10.1007/11744085_38

[78] Mikolajczyk, K., Schmid, C., “Scale & affine invariant interest point detectors”, International Journal of Computer Vision, Vol. 60, No. 1, 2004, str. 63–86, dostupno na: http://dx.doi.org/10.1023/B:VISI.0000027790.02288.f2

[79] Harris, C., Stephens, M., “A combined corner and edge detector”, in Proceedings of the Alvey Vision Conference, AVC 1988, Manchester, UK, September, 1988, 1988, str. 1–6, dostupno na: http://dx.doi.org/10.5244/C.2.23

[80] Matas, J., Chum, O., Urban, M., Pajdla, T., “Robust wide-baseline stereo from maximally stable extremal regions”, Image Vision Comput., Vol. 22, No. 10, 2004, str. 761–767, dostupno na: http://dx.doi.org/10.1016/j.imavis.2004.02.006

[81] Bay, H., Tuytelaars, T., Gool, L. J. V., “SURF: speeded up robust features”, in Computer Vision - ECCV 2006, 9th European Conference on Computer Vision, Graz, Austria, May 7-13, 2006, Proceedings, Part I, 2006, str. 404–417, dostupno na: http://dx.doi.org/10.1007/11744023_32

[82] Ojala, T., Pietikäinen, M., Mäenpää, T., “Multiresolution gray-scale and rotation invariant texture classification with local binary patterns”, IEEE Trans. Pattern Anal. Mach. Intell., Vol. 24, No. 7, 2002, str. 971–987, dostupno na: http://dx.doi.org/10.1109/TPAMI.2002.1017623

[83] Yosinski, J., Clune, J., Bengio, Y., Lipson, H., “How transferable are features in deep neural networks?”, in Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada, 2014, str. 3320–3328, dostupno na: http://papers.nips.cc/paper/5347-how-transferable-are-features-in-deep-neural-networks

[84] Chatfield, K., Simonyan, K., Vedaldi, A., Zisserman, A., “Return of the devil in the details: Delving deep into convolutional nets”, in British Machine Vision Conference, BMVC 2014, Nottingham, UK, September 1-5, 2014, 2014, dostupno na: http://www.bmva.org/bmvc/2014/papers/paper054/index.html

[85] Oquab, M., Bottou, L., Laptev, I., Sivic, J., “Learning and transferring mid-level image representations using convolutional neural networks”, in 2014 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2014, Columbus, OH, USA, June 23-28, 2014, 2014, str. 1717–1724, dostupno na: http://dx.doi.org/10.1109/CVPR.2014.222

[86] Razavian, A. S., Azizpour, H., Sullivan, J., Carlsson, S., “CNN features off-the-shelf: An astounding baseline for recognition”, in IEEE Conference on Computer Vision and Pattern Recognition, CVPR Workshops 2014, Columbus, OH, USA, June 23-28, 2014, 2014, str. 512–519, dostupno na: http://dx.doi.org/10.1109/CVPRW.2014.131

[87] Gordo, A., Gaidon, A., Perronnin, F., "Deep fishing: Gradient features from deep nets", in Proceedings of the British Machine Vision Conference 2015, BMVC 2015, Swansea, UK, September 7-10, 2015, 2015, str. 111.1–111.12, dostupno na: http://dx.doi.org/10.5244/C.29.111

[88] Bosch, A., Zisserman, A., Muñoz, X., "Image classification using random forests and ferns", in IEEE 11th International Conference on Computer Vision, ICCV 2007, Rio de Janeiro, Brazil, October 14-20, 2007, 2007, str. 1–8, dostupno na: http://dx.doi.org/10.1109/ICCV.2007.4409066

[89] Brown, M., Lowe, D. G., "Automatic panoramic image stitching using invariant features", International Journal of Computer Vision, Vol. 74, No. 1, 2007, str. 59–73, dostupno na: http://dx.doi.org/10.1007/s11263-006-0002-3

[90] Girshick, R. B., Donahue, J., Darrell, T., Malik, J., "Rich feature hierarchies for accurate object detection and semantic segmentation", in 2014 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2014, Columbus, OH, USA, June 23-28, 2014, 2014, str. 580–587, dostupno na: http://dx.doi.org/10.1109/CVPR.2014.81

[91] Hariharan, B., Arbeláez, P. A., Girshick, R. B., Malik, J., "Simultaneous detection and segmentation", in Computer Vision - ECCV 2014 - 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part VII, 2014, str. 297–312, dostupno na: http://dx.doi.org/10.1007/978-3-319-10584-0_20

[92] Long, J., Shelhamer, E., Darrell, T., "Fully convolutional networks for semantic segmentation", in IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015, 2015, str. 3431–3440, dostupno na: http://dx.doi.org/10.1109/CVPR.2015.7298965

[93] Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., Fei-Fei, L., "ImageNet: A Large-Scale Hierarchical Image Database", in CVPR09, 2009.

[94] Krapac, J., Šegvić, S., "Predavanja iz predmeta Duboko učenje", dostupno na: http://www.zemris.fer.hr/~ssegvic/du/ (8. prosinca 2016.).

[95] He, K., Zhang, X., Ren, S., Sun, J., "Deep residual learning for image recognition", in 2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016, 2016, str. 770–778, dostupno na: https://doi.org/10.1109/CVPR.2016.90

[96] Huang, G., Liu, Z., Weinberger, K. Q., "Densely connected convolutional networks", CoRR, Vol. abs/1608.06993, 2016, dostupno na: http://arxiv.org/abs/1608.06993

[97] Chatfield, K., Lempitsky, V. S., Vedaldi, A., Zisserman, A., "The devil is in the details: an evaluation of recent feature encoding methods", in British Machine Vision Conference, BMVC 2011, Dundee, UK, August 29 - September 2, 2011. Proceedings, 2011, str. 1–12, dostupno na: http://dx.doi.org/10.5244/C.25.76

[98] Ng, A., Duchi, J., "Predavanja iz predmeta Strojno učenje", dostupno na: http://cs229.stanford.edu/ (04. siječnja 2017.).

[99] Theodoridis, S., Koutroumbas, K., Pattern Recognition, Fourth Edition, 4th ed. Academic Press, 2008.

[100] Gyrgyek, L., Pavešić, N., Ribarić, S., Uvod u raspoznavanje uzoraka: primjena računala i mikroračunala u sustavima za raspoznavanje uzoraka, 1st ed. Tehnička knjiga, 1998.

[101] Baecchi, C., Turchini, F., Seidenari, L., Bagdanov, A. D., Bimbo, A. D., "Fisher vectors over random density forests for object recognition", in 22nd International Conference on Pattern Recognition, ICPR 2014, Stockholm, Sweden, August 24-28, 2014, 2014, str. 4328–4333, dostupno na: http://dx.doi.org/10.1109/ICPR.2014.712

[102] Vedaldi, A., Fulkerson, B., "VLFeat: An open and portable library of computer vision algorithms", http://www.vlfeat.org/, 2008.

[103] Lloyd, S. P., "Least squares quantization in PCM", IEEE Trans. Information Theory, Vol. 28, No. 2, 1982, str. 129–136, dostupno na: http://dx.doi.org/10.1109/TIT.1982.1056489

[104] Elkan, C., "Using the triangle inequality to accelerate k-means", in Machine Learning, Proceedings of the Twentieth International Conference (ICML 2003), August 21-24, 2003, Washington, DC, USA, 2003, str. 147–153, dostupno na: http://www.aaai.org/Library/ICML/2003/icml03-022.php

[105] Beis, J. S., Lowe, D. G., "Shape indexing using approximate nearest-neighbour search in high-dimensional spaces", in 1997 Conference on Computer Vision and Pattern Recognition (CVPR'97), June 17-19, 1997, San Juan, Puerto Rico, 1997, str. 1000–1006, dostupno na: http://dx.doi.org/10.1109/CVPR.1997.609451

[106] Silpa-Anan, C., Hartley, R. I., "Optimised KD-trees for fast image descriptor matching", in 2008 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2008), 24-26 June 2008, Anchorage, Alaska, USA, 2008, dostupno na: http://dx.doi.org/10.1109/CVPR.2008.4587638

[107] Titterington, D., Smith, A. F. M., Makov, U. E., Statistical analysis of finite mixture distributions, ser. Wiley series in probability and mathematical statistics. Chichester: J. Wiley, 1985.

[108] Krapac, J., Verbeek, J. J., Jurie, F., "Modeling spatial layout with Fisher vectors for image categorization", in IEEE International Conference on Computer Vision, ICCV 2011, Barcelona, Spain, November 6-13, 2011, 2011, str. 1487–1494, dostupno na: http://dx.doi.org/10.1109/ICCV.2011.6126406

[109] Riesenhuber, M., Poggio, T., "Hierarchical models of object recognition in cortex", Nature Neuroscience, Vol. 2, 1999, str. 1019–1025.

[110] Wang, J., Yang, J., Yu, K., Lv, F., Huang, T. S., Gong, Y., "Locality-constrained linear coding for image classification", in The Twenty-Third IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2010, San Francisco, CA, USA, 13-18 June 2010, 2010, str. 3360–3367, dostupno na: http://dx.doi.org/10.1109/CVPR.2010.5540018

[111] Philbin, J., Chum, O., Isard, M., Sivic, J., Zisserman, A., "Lost in quantization: Improving particular object retrieval in large scale image databases", in 2008 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2008), 24-26 June 2008, Anchorage, Alaska, USA, 2008, dostupno na: http://dx.doi.org/10.1109/CVPR.2008.4587635

[112] Drmač, Z., Hari, V., Marušić, M., Rogina, M., Singer, S., Singer, S., Numerička analiza. Zagreb: Springer: predavanja i vježbe, 2003.

[113] Jégou, H., Perronnin, F., Douze, M., Sánchez, J., Pérez, P., Schmid, C., "Aggregating local image descriptors into compact codes", IEEE Trans. Pattern Anal. Mach. Intell., Vol. 34, No. 9, 2012, str. 1704–1716, dostupno na: http://dx.doi.org/10.1109/TPAMI.2011.235

[114] Liu, D., Hua, G., Viola, P. A., Chen, T., "Integrated feature selection and higher-order spatial feature extraction for object categorization", in 2008 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2008), 24-26 June 2008, Anchorage, Alaska, USA, 2008, dostupno na: http://dx.doi.org/10.1109/CVPR.2008.4587403

[115] Savarese, S., Winn, J. M., Criminisi, A., "Discriminative object class models of appearance and shape by correlatons", in 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2006), 17-22 June 2006, New York, NY, USA, 2006, str. 2033–2040, dostupno na: http://dx.doi.org/10.1109/CVPR.2006.102

[116] Yang, J., Yu, K., Huang, T. S., "Efficient highly over-complete sparse coding using a mixture model", in Computer Vision - ECCV 2010 - 11th European Conference on Computer Vision, Heraklion, Crete, Greece, September 5-11, 2010, Proceedings, Part V, 2010, str. 113–126, dostupno na: http://dx.doi.org/10.1007/978-3-642-15555-0_9

[117] Sánchez, J., Perronnin, F., de Campos, T. E., "Modeling the spatial layout of images beyond spatial pyramids", Pattern Recognition Letters, Vol. 33, No. 16, 2012, str. 2216–2223, dostupno na: http://dx.doi.org/10.1016/j.patrec.2012.07.019

[118] Sharma, G., Jurie, F., "Learning discriminative spatial representation for image classification", in British Machine Vision Conference, BMVC 2011, Dundee, UK, August 29 - September 2, 2011. Proceedings, 2011, str. 1–11, dostupno na: http://dx.doi.org/10.5244/C.25.6

[119] Bosch, A., Zisserman, A., Muñoz, X., "Representing shape with a spatial pyramid kernel", in Proceedings of the 6th ACM International Conference on Image and Video Retrieval, CIVR 2007, Amsterdam, The Netherlands, July 9-11, 2007, 2007, str. 401–408, dostupno na: http://doi.acm.org/10.1145/1282280.1282340

[120] Agarwal, A., Triggs, B., "Hyperfeatures - multilevel local coding for visual recognition", in Computer Vision - ECCV 2006, 9th European Conference on Computer Vision, Graz, Austria, May 7-13, 2006, Proceedings, Part I, 2006, str. 30–43, dostupno na: http://dx.doi.org/10.1007/11744023_3

[121] Perronnin, F., "Universal and adapted vocabularies for generic visual categorization", IEEE Trans. Pattern Anal. Mach. Intell., Vol. 30, No. 7, 2008, str. 1243–1256, dostupno na: http://dx.doi.org/10.1109/TPAMI.2007.70755

[122] Yang, Y., Newsam, S. D., "Spatial pyramid co-occurrence for image classification", in IEEE International Conference on Computer Vision, ICCV 2011, Barcelona, Spain, November 6-13, 2011, 2011, str. 1465–1472, dostupno na: http://dx.doi.org/10.1109/ICCV.2011.6126403

[123] Singh, S., Gupta, A., Efros, A. A., "Unsupervised discovery of mid-level discriminative patches", in Computer Vision - ECCV 2012 - 12th European Conference on Computer Vision, Florence, Italy, October 7-13, 2012, Proceedings, Part II, 2012, str. 73–86, dostupno na: http://dx.doi.org/10.1007/978-3-642-33709-3_6

[124] Šmuc, T., "Predavanja iz predmeta Strojno učenje", dostupno na: https://web.math.pmf.unizg.hr/nastava/su/materijali/ (1. ožujka 2017.).

[125] Fernando, B., Fromont, É., Tuytelaars, T., "Mining mid-level features for image classification", International Journal of Computer Vision, Vol. 108, No. 3, 2014, str. 186–203, dostupno na: http://dx.doi.org/10.1007/s11263-014-0700-1

[126] Weng, C., Yuan, J., "Efficient mining of optimal AND/OR patterns for visual recognition", IEEE Trans. Multimedia, Vol. 17, No. 5, 2015, str. 626–635, dostupno na: http://dx.doi.org/10.1109/TMM.2015.2414720

[127] Russakovsky, O., Lin, Y., Yu, K., Li, F., "Object-centric spatial pooling for image classification", in Computer Vision - ECCV 2012 - 12th European Conference on Computer Vision, Florence, Italy, October 7-13, 2012, Proceedings, Part II, 2012, str. 1–15, dostupno na: http://dx.doi.org/10.1007/978-3-642-33709-3_1

[128] Pandey, M., Lazebnik, S., "Scene recognition and weakly supervised object localization with deformable part-based models", in IEEE International Conference on Computer Vision, ICCV 2011, Barcelona, Spain, November 6-13, 2011, 2011, str. 1307–1314, dostupno na: http://dx.doi.org/10.1109/ICCV.2011.6126383

[129] Yuan, J., Wu, Y., Yang, M., "Discovery of collocation patterns: from visual words to visual phrases", in 2007 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2007), 18-23 June 2007, Minneapolis, Minnesota, USA, 2007, dostupno na: http://dx.doi.org/10.1109/CVPR.2007.383222

[130] Torralba, A., Murphy, K. P., Freeman, W. T., "Sharing visual features for multiclass and multiview object detection", IEEE Trans. Pattern Anal. Mach. Intell., Vol. 29, No. 5, 2007, str. 854–869, dostupno na: http://dx.doi.org/10.1109/TPAMI.2007.1055

[131] Everingham, M., Gool, L., Williams, C. K., Winn, J., Zisserman, A., "The Pascal visual object classes (VOC) challenge", Int. J. Comput. Vision, Vol. 88, No. 2, Jun. 2010, str. 303–338.

[132] Lampert, C. H., Blaschko, M. B., Hofmann, T., "Efficient subwindow search: A branch and bound framework for object localization", IEEE Trans. Pattern Anal. Mach. Intell., Vol. 31, No. 12, 2009, str. 2129–2142, dostupno na: http://dx.doi.org/10.1109/TPAMI.2009.144

[133] Krapac, J., Šegvić, S., "Weakly-supervised semantic segmentation by redistributing region scores back to the pixels", in Pattern Recognition - 38th German Conference, GCPR 2016, Hannover, Germany, September 12-15, 2016, Proceedings, 2016, str. 377–388, dostupno na: http://dx.doi.org/10.1007/978-3-319-45886-1_31

[134] Alexe, B., Deselaers, T., Ferrari, V., “Measuring the objectness of image windows”, IEEE Trans. Pattern Anal. Mach. Intell., Vol. 34, No. 11, 2012, str. 2189–2202, dostupno na: http://dx.doi.org/10.1109/TPAMI.2012.28

[135] Viola, P., Jones, M., “Robust real-time object detection”, International Journal of Computer Vision, Vol. 4, 2001.

[136] van de Sande, K. E. A., Snoek, C. G. M., Smeulders, A. W. M., “Fisher and VLAD with FLAIR”, in 2014 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2014, Columbus, OH, USA, June 23-28, 2014, 2014, str. 2377–2384, dostupno na: http://dx.doi.org/10.1109/CVPR.2014.304

[137] Tolias, G., Sicre, R., Jégou, H., “Particular object retrieval with integral max-pooling of CNN activations”, CoRR, Vol. abs/1511.05879, 2015, dostupno na: http://arxiv.org/abs/1511.05879

[138] Huang, C., Ai, H., Li, Y., Lao, S., “High-performance rotation invariant multiview face detection”, IEEE Trans. Pattern Anal. Mach. Intell., Vol. 29, No. 4, 2007, str. 671–686, dostupno na: http://dx.doi.org/10.1109/TPAMI.2007.1011

[139] Wu, B., Nevatia, R., “Cluster boosted tree classifier for multi-view, multi-pose object detection”, in IEEE 11th International Conference on Computer Vision, ICCV 2007, Rio de Janeiro, Brazil, October 14-20, 2007, 2007, str. 1–8, dostupno na: http://dx.doi.org/10.1109/ICCV.2007.4409006

[140] Cortes, C., Vapnik, V., “Support-vector networks”, Machine Learning, Vol. 20, No. 3, 1995, str. 273–297, dostupno na: http://dx.doi.org/10.1007/BF00994018

[141] Felzenszwalb, P. F., Huttenlocher, D. P., “Efficient graph-based image segmentation”, International Journal of Computer Vision, Vol. 59, No. 2, 2004, str. 167–181, dostupno na: http://dx.doi.org/10.1023/B:VISI.0000022288.19776.77

[142] Javor, P., Matematička analiza 2. Zagreb, Hrvatska: Element, 2002.

[143] Rabinovich, A., Belongie, S. J., Lange, T., Buhmann, J. M., “Model order selection and cue combination for image segmentation”, in 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2006), 17-22 June 2006, New York, NY, USA, 2006, str. 1130–1137, dostupno na: http://dx.doi.org/10.1109/CVPR.2006.186

[144] Zhou, B., Khosla, A., Lapedriza, À., Oliva, A., Torralba, A., “Learning deep features for discriminative localization”, in 2016 IEEE Conference on Computer Vision and

Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016, 2016, str. 2921–2929, dostupno na: http://dx.doi.org/10.1109/CVPR.2016.319

[145] Bency, A. J., Kwon, H., Lee, H., Karthikeyan, S., Manjunath, B. S., "Weakly supervised localization using deep feature maps", in Computer Vision - ECCV 2016 - 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part I, 2016, str. 714–731, dostupno na: http://dx.doi.org/10.1007/978-3-319-46448-0_43

[146] Siva, P., Russell, C., Xiang, T., "In defence of negative mining for annotating weakly labelled data", in Computer Vision - ECCV 2012 - 12th European Conference on Computer Vision, Florence, Italy, October 7-13, 2012, Proceedings, Part III, 2012, str. 594–608, dostupno na: http://dx.doi.org/10.1007/978-3-642-33712-3_43

[147] Alexe, B., Deselaers, T., Ferrari, V., "What is an object?", in The Twenty-Third IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2010, San Francisco, CA, USA, 13-18 June 2010, 2010, str. 73–80, dostupno na: http://dx.doi.org/10.1109/CVPR.2010.5540226

[148] Ke, Y., Tang, X., Jing, F., "The design of high-level features for photo quality assessment", in 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2006), 17-22 June 2006, New York, NY, USA, 2006, str. 419–426, dostupno na: http://dx.doi.org/10.1109/CVPR.2006.303

[149] Vedaldi, A., Zisserman, A., "Efficient additive kernels via explicit feature maps", IEEE Trans. Pattern Anal. Mach. Intell., Vol. 34, No. 3, 2012, str. 480–492, dostupno na: http://dx.doi.org/10.1109/TPAMI.2011.153

[150] Novotný, D., Larlus, D., Perronnin, F., Vedaldi, A., "Understanding the Fisher Vector: a multimodal part model", CoRR, Vol. abs/1504.04763, 2015, dostupno na: http://arxiv.org/abs/1504.04763

[151] Cinbis, R. G., Verbeek, J. J., Schmid, C., "Image categorization using fisher kernels of non-iid image models", in 2012 IEEE Conference on Computer Vision and Pattern Recognition, Providence, RI, USA, June 16-21, 2012, 2012, str. 2184–2191, dostupno na: http://dx.doi.org/10.1109/CVPR.2012.6247926

[152] Perronnin, F., Liu, Y., Sánchez, J., Poirier, H., "Large-scale image retrieval with compressed fisher vectors", in The Twenty-Third IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2010, San Francisco, CA, USA, 13-18 June 2010, 2010, str. 3384–3391, dostupno na: http://dx.doi.org/10.1109/CVPR.2010.5540009

[153] Fawcett, T., “An introduction to ROC analysis”, Pattern Recognition Letters, Vol. 27, No. 8, 2006, str. 861–874, dostupno na: http://dx.doi.org/10.1016/j.patrec.2005.10.010

[154] Inland transport comitee, “Convention on road signs and signals”, Economic comission for Europe, 1968.

[155] Douze, M., Jégou, H., “The Yael library”, in Proceedings of the ACM International Conference on Multimedia, MM’14, Orlando, FL, USA, November 03 - 07, 2014, 2014, str. 687–690, dostupno na: http://doi.acm.org/10.1145/2647868.2654892

[156] Mairal, J., Bach, F. R., Ponce, J., “Sparse modeling for image and vision processing”, Foundations and Trends in Computer Graphics and Vision, Vol. 8, No. 2-3, 2014, str. 85–283, dostupno na: http://dx.doi.org/10.1561/0600000058

[157] Itseez, “Open source computer vision library”, https://github.com/itseez/opencv, 2017.

[158] Krapac, J., Šegvić, S., “Weakly supervised object localization with large fisher vectors”, in VISAPP 2015 - Proceedings of the 10th International Conference on Computer Vision Theory and Applications, Volume 2, Berlin, Germany, 11-14 March, 2015., 2015, str. 44–53, dostupno na: http://dx.doi.org/10.5220/0005294900440053

[159] OpenStreetMap Wiki, “Video mapping — OpenStreetMap Wiki,”, dostupno na: http://wiki.openstreetmap.org/w/index.php?title=Video_mapping&oldid=1316747 (24. prosinca 2016.).

[160] Ramm, F., Topf, J., Chilton, S., OpenStreetMap: Using and Enhancing the Free Map of the World. Uit Cambridge Limited, 2010, dostupno na: http://books.google.de/books?id=AnCNQQAACAAJ

[161] WGS 84 Implementation Manual, Brussels, Belgium, 1998, dostupno na: http://www.icao.int/safety/pbn/Documentation/EUROCONTROL/Eurocontrol%20WGS%2084%20Implementation%20Manual.pdf

[162] Geofabrik, “OpenStreetMap Data Extracts”, dostupno na: http://http://download.geofabrik.de/ (24. prosinca 2016.).

[163] OpenStreetMap Wiki, “Overpass API — OpenStreetMap Wiki,”, dostupno na: http://wiki.openstreetmap.org/wiki/Overpass_API (24. prosinca 2016.).

[164] Peručić, F., Sustav za kartiranje elemenata sigurnosti cestovne infrastrukture pomoću georeferenciranog videa, diplomski rad, Geodetski fakultet, Sveučilište u Zagrebu, Zagreb, 2013.

[165] Beck, A., Teboulle, M., “A fast iterative shrinkage-thresholding algorithm for linear inverse problems”, SIAM J. Imaging Sciences, Vol. 2, No. 1, 2009, str. 183–202, dostupno na: http://dx.doi.org/10.1137/080716542

[166] Rossum, G., Python Reference Manual, Amsterdam, The Netherlands, The Netherlands, 1995.

[167] Krapac, J., Šegvić, S., “Fast approximate GMM soft-assign for fine-grained image classification with large fisher vectors”, in Pattern Recognition - 37th German Conference, GCPR 2015, Aachen, Germany, October 7-10, 2015, Proceedings, 2015, str. 470–480, dostupno na: http://dx.doi.org/10.1007/978-3-319-24947-6_39

Popis slika

Popis tablica

Životopis

Valentina Zadrija rođena je 27. veljače 1985. godine u Zagrebu, a prirodoslovno-matematičku XV. gimnaziju završila je 2003. godine u Zagrebu. Iste godine upisala je diplomski studj računarstva na Fakultetu elektrotehnike i računarstva Sveučilišta u Zagrebu. Diplomirala je 2008. godine pod vodstvom prof. dr. sc. Slobodana Ribarića s temom "Postupci zaključivanja za shemu za predstavljanje znanja temeljenu na neizrazitim Petrijevim mrežama".

Od listopada 2008. godine do lipnja 2009. bila je zaposlena kao zavodski suradnik na Zavodu za elektroniku, mikroelektroniku, računalne i inteligentne sustave Fakulteta elektrotehnike i računarstva u Zagrebu, a od lipnja 2009. godine do kolovoza 2010. godine kao znanstveni novak na istome zavodu. Bila je uključena u nastavne aktivnosti Zavoda na predmetima Napredni operacijski sustavi, Analiza i projektiranje računalom, Memorijski sustavi i Pouzdanost računalnih sustava, a asistirala je u vođenju diplomskih i završnih radova. Od kolovoza 2010. godine do danas zaposlena je kao razvojni inženjer programskih sustava u poduzeću Mireo d.d. gdje se bavi digitalnom kartografijom.

U okviru poslijediplomskog studija, sudjelovala je na više domaćih i inozemnih znanstvenih projekata, uključujući bilateralni austrijsko-hrvatski projekt "Kartiranje i verifikacija prometne signalizacije". Njezini istraživački interesi uključuju područja računalnog vida i strojnog učenja. U suautorstvu je objavila pet radova na međunarodnim znanstvenim skupovima. Dobro poznaje engleski jezik, a služi se i njemačkim jezikom. Udana je i ima troje djece.

Popis objavljenih djela

1. Zadrija, V. Krapac, J., Verbeek, J., Šegvić, S., „Patch-level Spatial Layout for Classification and Weakly Supervised Localization”, 37th German Conference on Pattern Recognition Aachen, Njemačka, 2015
2. Zadrija, V. Šegvić, S., „Experimental Evaluation of Multiplicative Kernel SVM Classifiers for Multi-Class Detection”, Third Croatian Computer Vision Workshop (CCVW 2014) Zagreb, Hrvatska, 2014
3. Zadrija, V. Šegvić, S., „Multiclass Road Sign Detection using Multiplicative Kernel”, Second Croatian Computer Vision Workshop (CCVW 2013) Zagreb, Hrvatska, 2013

4. Ribarić, S., Zadrija, V., „An Object-Oriented Implementation of a Knowledge Representation Scheme based on Fuzzy Petri Nets", Seventh Int. Conference on Fuzzy Systems and Knowledge Discovery, Yantai, Kina, 2010
5. Ribarić, S. Pavešić, N., Zadrija, V., „Intersection Search for a Fuzzy Petri Net-Based Knowledge Representation Scheme", 13th International Conference on Knowledge-Based and Intelligent Information and Engineering Systems Santiago, Čile, 2009

Biography

Valentina Zadrija was born on 27th of February 1985 in Zagreb, Croatia. She completed her high school education in 2003, by graduating from Fifteenth Mathematical Gymnasium in Zagreb. She received her BSc degree in computer science from the Faculty of Electrical Engineering and Computing (FER) in 2008, under the expert guidance of Slobodan Ribarić, PhD. The title of her BSc thesis was "Inference procedures for knowledge representation scheme based on fuzzy Petri Nets".

From October 2008 to June 2009, she was employed at the Department of Electronics, Microlelectronics, Computer and Intelligent Systems at FER, as a research associate. From June 2009 to August 2010, she worked at the same Faculty as a research assistant. She was a teaching assistant on several courses from FER, such as Advanced Operating Systems, Computer Aided Analysis and Design, Memory Systems and Computer Systems Reliability. She also supervised several BSc and MSc student theses. Since August 2010, she has been employed as a software engineer at Mireo PLC, where her professional interests include digital cartography.

Valentina Zadrija participated in several research projects during her postgraduate study, such as international project "Mapping and assessing traffic infrastructure" co-funded by TU Graz, Austria. Her research interests include computer vision and machine learning. She has co-authored five papers published in international conferences. She is fluent in english, and has basic communication skills in german. She is married and has three children.