Enhanced Semantic Descriptors for Functional Scene Categorization

Gloria Zen[1], Negar Rostamzadeh[1], Jacopo Staiano[1], Elisa Ricci[2], Nicu Sebe[1]
[1]*DISI, University of Trento, Italy.* [2]*DIEI, University of Perugia, Italy.*
{*zen, rostamzadeh, staiano, sebe*}*@disi.unitn.it, elisa.ricci@diei.unipg.it*

Abstract

In this work we present a novel approach which combines semantic information with low level features extracted from a complex video scene. The proposed method for video scene understanding relies on a bag-of-words approach, in which, typically, visual words contain information of local motion, but information regarding what generated such motion is discarded. Instead, in our framework, the semantic information is embedded in the visual words and it allows to automatically obtain semantic categorization of the scene. We show the effectiveness of our method in a traffic analysis scenario: in this case, two main semantic classes, pedestrians and vehicles, are discovered.

1. Introduction

Recently, non-object-centric approaches for the analysis of dynamic scenes have gained popularity and have been proven effective for many application scenarios, *e.g.* extraction of salient activities [20], scene semantic segmentation [4], *etc*. In particular, in the case of complex scenarios with many targets and occlusions, they are preferred to the classic object detection/tracking schema because they are not subject to the problem of broken trajectories or to the curse of dimensionality when trying to consider the spatio-temporal correlation between many targets. In a nutshell, non-object-centric methods rely on low level cues (*e.g.* local motion and foreground) that are analyzed in a bag-of-words framework. Firstly, a visual codebook of the scene is generated, where visual words generally encode information about the local motion in the scene. Secondly, the video is divided into (i) short clips (or spatial patches) and for each of them a bag-of-visual-words is build. Starting from this bag-of-words representation of a video stream, statistical models such as Probabilistic Topic Models (PTM) are deployed in order to mine (i), the typical patterns of behavior and the anomalies from the scene or (ii), the spatial segmentation of the scene, where patches with similar motion behavior are assigned to the same semantic area. This approach has shown to be robust to noise; however, there are some limitations. Switching from an object-centric to a non-object-centric perspective, the method gains in robustness w.r.t noise and to broken trajectories; however, at the same time, the information about the semantic of the objects causing the detected motion (*e.g.* cars, pedestrians, *etc.*) is discarded. As a consequence, it is difficult to reason about topics extracted as they do not always correspond to high-level description of the scene according to a human observer. In this work, we focus on this aspect and propose the use of enhanced semantic descriptors, as a step towards filling the semantic gap between object-centric and non-object-centric approaches. We show our results on publicly available datasets and compare them with recent related work.

2. Related Work

In the last decades, the bag-of-words paradigm has been widely adopted firstly in still images analysis [2, 3] and then in dynamic scenes analysis [19, 8, 6]. This paradigm is composed of two main steps: (i) codebook generation (ii) bag-of-words formation and clustering. In detail, the first phase corresponds to the definition of the scene descriptors and thus it strongly influences the final results and the effectiveness of the method. In other words, if the features we extract from our scene do not properly represent the activities occurring in the scene, or some important information is discarded in this phase, the topics extracted during the second phase may not be a representative synopsis of the scene observed. In particular, by discarding semantic information in this preliminary step of the analysis, the final topics extracted may not correspond to the high level segmentation generated by a human observed. This problem has been addressed by [5] with Object Bank for still images analysis: the intuition is that scene descriptors are combined with information of the objects

in the scene. While obtaining state-of-the-art performance, this method is found to be very expensive when a large number of categories is considered (as in the case of still images). In the case of video scene analysis, much effort has been devoted to enrich the descriptors with additional information beyond motion. Messing *et al.* [6] used descriptors which encode information of both local motion and appearance, in order to distinguish between actions with similar motion but different appearance (*e.g.* eating snack from eating banana) and between actions with similar appearance but different motion (*e.g.* peeling banana from eating banana). Additionally, they used sparse information provided by a face detector to augment their features with relative position information. Shitrit *et al.* [12] track people by linking sparse information of people detection into tracklets. This approach is proven to be more effective in term of robusteness and complexity then recursively track from frame to frame. Sangmin *et al.* [9] detected functional objects in traffic scenes (*e.g.* delivery track) by using features which encode relation and actions w.r.t the scene context. Chen *et al.* [1] proposed the use of motion features (MoSIFT) to encode static image appearance features together with motion information.

Based on [16], Papageorgiu *et. al* [10] proposed a general object detection scheme using Haar wavelets and SVMs and applied it for face, pedestrian and car detection. Other works focused specifically on pedestrian [17] or car [21] detection: Viola *et al.* [17] were able to detect pedestrians at very small scales (up to 20×15 pixels) by using both appearance and motion information; Zhu *et al.* [21] devised a car detection scheme exploiting global structure and local texture features. In order to reduce human efforts in the labeling task, methods based on semi-supervision or active learning have been proposed [11, 7]. Finally, a relevant work [18] proposed a method for adapting a generic pedestrian detector to specific traffic scenes, exploiting multiple cues such as size and motion.

3. Our method

Similarly to Turek *et al.* [15], our aim is to perform a semantic segmentation of the scene, where the scene element categories are primarily defined by their behavior, rather then their appearance or shape. In particular, Turek *et al.* [15] use a hierarchy of motion features. They divide the scene into several patches and for each patch a codebook histogram is formed. Patches are then clustered according to their histograms' similarity and the patches belonging to the same cluster are associated to the same functional category. Differently from them, in our work the semantic information of the entity causing the motion is embedded in the descriptor. This is obtained by combining information provided by the pedestrian and vehicle detectors with local motion information.

For the pedestrians and cars detector we rely on [17]. For low level cues extraction, we use Lucas Kanade algorithm [14] for optical flow combined with dynamic Gaussian-Mixture background model [13]. Motion vectors are quantized into 8 possible directions while patches with only static points and sufficient foreground are considered associated to a static event (*e.g* pedestrian stopped at red traffic light waiting to cross the street). Then, for each patch we build 2 histograms of 9 bins each (8 bins identify motion and 1 bin is for static events).

In details, our method is formulated as follows. We divide our scene into $N_x \times N_y$ patches. Then, for each of these patches $p_{i,j}$, where $i = 1, ..., N_y$ and $j = 1, ..., N_x$, we build 2 histograms of cars h_C and pedestrians events h_P by analyzing a video sequence of at least 30 minutes of length. In particular, we consider a couple of frames i and $i - I_s$ at a time, where I_s is the step for optical flow computation. For each frame i, we compute the corresponding foreground mask F^i and, still on frame i, we run the car and the pedestrian detectors. We define $B_P^i = \{b_1, ..., b_{N_P^i}\}$ and $B_C^i = \{b_1, ..., b_{N_C^i}\}$ as the resulting bounding boxes found, localizing respectively pedestrians (P) and cars (C), The bounding box $b_k = (x, y, w, h)$, with $k = 1, ..., N_K^i$ and $K = \{P, C\}$, is defined by the coordinate of its upper left corner (x, y) and its size (w, h). N_K^i is the number of items found, which varies at each frame i. Every time a static or a motion event is detected, it is counted as an occurrence in the corresponding h_P or h_C histogram, depending on if it is included, respectively, in one of the bounding box from B_P or B_C. Once the histograms of the patches are built, we cluster them in order to obtain the spatial segmentation of the scene. Similarly to [15], we label cells by using k-means and mean-shift. We obtain 4 spatial segmentations by considering for clustering:

- 18 bins *semantic* histograms, obtained by concatenating h_P and h_C
- 9 bins *non semantic* histograms, obtained by summing h_P with h_C,
- 9 bins *only-car semantic* histograms h_C
- 9 bins *only-pedestrian semantic* histograms h_P

The effect of differently combining h_P and h_C are discussed in the experimental session. Once the semantic maps of the scene are built, they can be used for further analyses of events in the scenes.

Figure 1. Dataset used. (Left) MIT Traffic and (Right) QMUL Junction dataset

Table 1. Details on experimental setup

	n° frames	fps	video duration	frame size $W \times H$	patch size $s_W \times s_H$	n° patches $N_x \times N_y$
MIT Traffic	162000	30	$1h$ 30'	720×480	15×15	48×32
QMUL Junction	90000	25	$1h$	360×288	12×12	30×24

It is worth noting that in application with crowded scene scenarios the object centric paradigm based on detection/tracking is not reliable because of the high risk of tracking failures and the curse of dimensionality. However, the use of sparse information provided by the object detectors combined with low level features allows to exploit the advantages of both non object centric and object centric methods; respectively these advantages include (i) robustness to noise and reduced computational complexity on one side and, on the other one, (ii) encoding semantic information of the entities interacting in the scene.
Another advantage of the method is that we can perform an analysis of the scene at two different levels of optical flow. This allows to detect objects at different speeds, like in the case of cars and pedestrians. In fact, by using a low value of I_s, the shift measured of a slow moving object (*e.g.*, a pedestrian) is close to zero and thus it is detected as a static event. On the other side, by increasing the value of I_s, we can detect the motion of slow moving object, but the motion cues extracted for high speed object tend to be very noisy. In our experiment we set $I_s^C = 5$ and $I_s^P = 10$.

4. Results

We show our results on two public datasets: MIT Traffic dataset[1] and QMUL Junction dataset[2]. Sample frames extracted from these two datasets are shown in Fig. 1; details on the datasets and the experimental setup are summarized in Table 1. Our semantic segmentation method's results are visually displayed in Fig. 2 and 3, respectively for the MIT Traffic and QMUL Junction datasets.

[1] www.ee.cuhk.edu.hk/~xgwang/MITtraffic.html
[2] www.eecs.qmul.ac.uk/~jianli/Junction.html

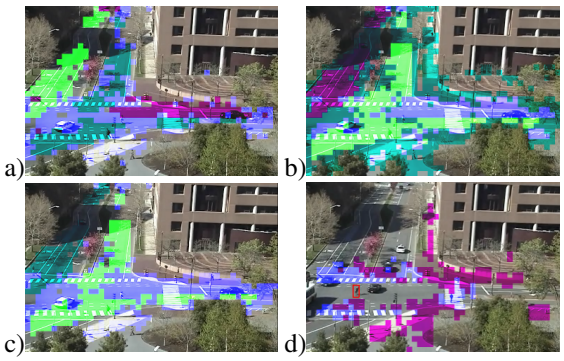

Figure 2. MIT dataset. Scene segmentation based on (a) semantic, (b) non-semantic, (c) car semantic and (d) pedestrian semantic descriptors.

4.1 MIT Traffic dataset

Visual inspection of our segmentation results on the MIT Traffic dataset (Fig. 2) against the path model reported and exploited in [18], shows that our proposed method performs comparably while being extremely simple and cheap to implement. It requires, in fact, the availability of a detector for the objects of interest in the scene and makes use of simple and well-known features, such as optical flow: for both the requirements, extensive literature and code are available. Figure 2(b) shows that with our method, by using only non semantic descriptors, some areas like zebra crossing are not distinguished. By using semantic descriptors the areas associated to pedestrians or cars are better distinguished, as it can be seen in Fig.2(a). Still, as observable from Fig.2(c) and Fig.2(d), clustering the two histograms separated may provide the best solutions. In (c) we distinguish between different cars lanes, associated to the different main motion directions, while in (d) we can distinguish between two different conceptual areas for pedestrians: zebra crossing (highlighted in blue) and sidewalks (in violet). In Fig.2(d) additionally we show an example of anomaly that can be detected by using the semantic pedestrian map generated, *i.e.* a pedestrian walking outside the pedestrian area (in a red frame).

4.2 QMUL Junction dataset

In Fig. 3 a sample case of a pedestrian detected in an unusual position w.r.t the pedestrian semantic map of the scene is depicted. Due to its highly cluttered background, we could not get a reliable pedestrian detector to work with this dataset. Still, relying on ground-truth knowledge on some pedestrian positions we are able to infer, thanks to semantic scene segmentation, anomalies in the pedestrian behavior (e.g., crossing outside of zebras). Interestingly, observing the semantic pedestrian

Figure 3. (Left) Anomaly. Pedestrian detected outside associated semantic map. (Right) corresponding Foreground mask.

map shown in Fig 3(a), we notice that the two discovered clusters correspond to two conceptually different areas: i) zebra crossing, shown in blue, and ii) the waiting areas at the intersection between sidewalks and zebra crossings when the pedestrian traffic light is on red (shown in red). Our intuition is that the algorithm proposed by [16] could benefit in terms of speed, to some extent, and, more significantly, of a reduction in false positives by discarding detection sub-windows containing less than a given amount of foreground (shown in Fig.3(b)). Additionally, the semantic maps obtained by using the ground truth on pedestrians could be used for the same goal of [18] (i.e., filtering out false positives from automatically gathered training data in a novel scene). In future works we plan to present an extension of such algorithm in this direction.

5. Conclusions

We proposed a novel method which relies on the classical bag-of-words model but instead of ignoring the semantic information as it is typically done in the literature we include it in our descriptors. We presented results on two traffic datasets where two semantic object categories (cars and pedestrians) are present. While proving their effectiveness on anomaly detection tasks (*e.g.* pedestrian on the car lane), semantic maps can be useful for other tasks as shown in literature (*e.g.* automatic selection of training set [18]). Further work will involve considering other scenarios and extending the current object detector to exploit semantic maps for improving performance.

References

[1] M.-Y. Chen, A. Hauptmann, and H. Li. Combining motion understanding and keyframe image analysis for broadcast video information extraction. *Evolutionary and Bio-Inspired Computation: Theory and Applications IV, SPIE Defense, Security, and Sensing*, 2010.

[2] G. Csurka, C. R. Dance, L. Fan, J. Willamowski, and C. Bray. Visual categorization with bags of keypoints. *Workshop on Statistical Learning in Computer Vision, ECCV*, 2004.

[3] L. Fei-Fei and P. Perona. Natural scene categorization. *CVPR*, 2005.

[4] J. Li, S. Gong, and T. Xiang. Scene segmentation for behaviour correlation. *ECCV*, 2008.

[5] L.-J. Li, H. Su, E. P. Xing, and L. Fei-Fei. Object bank: A high-level image representation for scene classification and semantic feature sparsification. *Neural Information Processing Systems (NIPS)*, 2011.

[6] R. Messing, C. Pal, and H. Kautz. Activity recognition using the velocity histories of tracked keypoints. *ICCV*, 2009.

[7] T. T. Nguyen, N. D. Binh, and H. Bischof. Efficient boosting-based active learning for specific object detection problems. *Int. Journal of Electrical, Computer, and Systems Engineering*, 3:2070–3813, 2009.

[8] J. C. Niebles, H. Wang, and L. Fei-Fei. Unsupervised learning of human action categories using spatial-temporal words. *International Journal of Computer Vision*, 79(3):299–318, 2008.

[9] S. Oh, A. Hoogs, M. Turek, and R. Collins. Content-based retrieval of functional objects in video using scene context. *ECCV*, 2010.

[10] C. Papageorgiou and T. Poggio. A trainable system for object detection. *Int. Journal of Computer Vision (IJCV)*, 38(1):15–33, 2000.

[11] C. Rosenberg, M. Hebert, and H. Schneiderman. Semi-supervised self-training of object detection models. *IEEE Workshop on Applications of Computer Vision*, 2005.

[12] H. B. Shitrit, J. Berclaz, F. Fleuret, and P. Fua. Tracking multiple people under global appearance constraints. *ICCV*, 2011.

[13] C. Stauffer and W. Grimson. Adaptive background mixture models for real-time tracking. *CVPR*, 1999.

[14] C. Tomasi and T. Kanade. Detection and tracking of point features. *Technical Report CMU-CS-91-132, Carnegie Mellon University*, 1991.

[15] M. Turek, A. Hoogs, and R. Collins. Unsupervised learning of functional categories in video scenes. *ECCV*, 2010.

[16] P. Viola and M. Jones. Rapid object detection using a boosted cascade of simple features. *CVPR*, 2001.

[17] P. Viola, M. Jones, and D. Snow. Detecting pedestrians using patterns of motion and appearance. *ICCV*, 2003.

[18] M. Wang and X. Wang. Automatic adaptation of a generic pedestrian detector to a specific traffic scene. *CVPR*, 2011.

[19] X. Wang, X. Ma, and E. Grimson. Unsupervised activity perception in crowded and complicated scenes using hierarchical bayesian models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2009.

[20] G. Zen and E. Ricci. Earth mover's prototypes: A convex learning approach for discovering activity patterns in dynamic scenes. *CVPR*, 2011.

[21] Z. Zhu, H. Lu, J. Hu, and K. Uchimura. Car detection based on multi-cues integration. *International Conference of Pattern Recognition (ICPR)*, 2004.