Non-sequential Multi-view Detection, Localization and Identification of People Using Multi-modal Feature Maps

Rok Mandeljc, Stanislav Kovačič, Matej Kristan, Janez Perš

NOTE: this is a *pre-print version* of paper that was published in:
Computer Vision — ACCV 2012, Lecture Notes in Computer Science
Volume 7726, 2013, pp. 691-704

For *official version*, see:
http://dx.doi.org/10.1007/978-3-642-37431-9_53

Non-sequential Multi-view Detection, Localization and Identification of People Using Multi-modal Feature Maps

Rok Mandeljc, Stanislav Kovačič, Matej Kristan and Janez Perš

University of Ljubljana, Faculty of Electrical Engineering, Slovenia

Abstract. We present a novel multi-modal fusion framework for non-sequential person detection, localization and identification from multiple views. Our goal is independent processing of randomly-accessed sections of video, either individual frames or small batches thereof. This way, we aim to limit the error propagation that makes the existing approaches unsuitable for fully-autonomous tracking of multiple people in long video sequences. Our framework uses one or more trained classifiers to fuse multiple weak feature maps. We perform experimental validation on a challenging dataset, demonstrating how the framework can, depending on the provided feature maps, be used either only to improve generic person detection, or enable simultaneous detection and recognition of individuals. Finally, we show that tracking-by-identification using the output of the proposed framework outperforms the state-of-the-art identification-by-tracking approach in terms of preserved track identities.

1 Introduction

The need for unobtrusive recovery of individuals' *positions* and trajectories, measured in the *world coordinate system*, can be found in different scenarios, with most notable examples being closed-world surveillance applications, performance analysis in sports and sports medicine. Therefore, person detection and tracking using *multiple cameras with overlapping fields of view* is already a hot research topic. However, despite long tradition of multi-view multi-target tracking in computer vision, we are not aware of any tracker that would allow completely automatic, autonomous and unattended processing of very long video sequences involving multiple people in realistic indoor environments.

The existing approaches can be roughly divided into two groups. The first are so-called *detection-by-tracking* approaches, which are based on sequential techniques, such as Kalman filter or particle filters (e.g. [1,2] and [3,4], respectively). Such trackers are causal; they consider only information from previously-processed frames, which is why they are still considered state-of-the-art in the real-time tracking. However, relying on recursive tracking may result in irrecoverable errors when a person fails to be detected in a frame or when two detections made at different times are incorrectly linked. Errors tend to propagate and multiply in the subsequent frames. Eventually, such *unbounded error propagation* causes a tracker to fail and manual re-initialization is required.

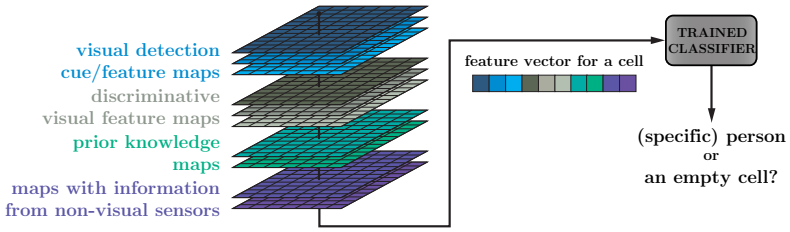


Fig. 1. Our framework uses multiple multi-modal features encoded as feature maps and fuses them using a trained classifier.

The other group, the *tracking-by-detection* approaches, employ robust frame-by-frame detection [5,6], on top of which global optimization methods are applied for tracking (e.g. [7]), usually off-line and in batch manner. However, when it comes to maintaining the identities of tracks, these approaches perform *identification-by-tracking*; they rely on identity propagation along the track, with none or limited appearance-based validation. As such, they are prone to *propagated identity switches* when people come close. Propagated identity switches manifest themselves as localization error after people disperse and, even if infrequent, make fully-automatic tracking over long time periods unfeasible. After a switch, the entire trajectory data is essentially invalid, and intervention of an operator is required. In absence of means to detect a propagation of an identity switch, the tracker needs to be constantly supervised.

We therefore aim to achieve non-sequential processing of individual frames or groups of few frames at a time. In terms of processing, this would enable a "random access" to the video sequence; based on previously-learned discriminative features, individuals' identities could be (re)established at any point, for example after they come together and disperse. Due to independent processing, the errors are not propagated to subsequent frames, and hence the error is always bounded. The obvious disadvantage of non-sequential identification is that it is much harder to achieve; to distinguish between individuals, a large amount of weak discriminative features might be required.

The main contribution of this paper is a novel multi-modal fusion framework for simultaneous person detection, localization and identification. We use multiple weak features encoded as feature maps, a generalization of the well-known concept of an occupancy map, and fuse the feature maps using one or more trained classifiers. In the paper, we also further the notion of *tracking-by-identification* as an alternative to the state of the art, which predominantly focuses on identification-by-tracking.

The remainder of the paper is structured as follows. In Section 2, an overview of related work is given, followed by the introduction of our framework in Section 3 and details on feature maps in Section 4. Section 5 describes experimental validation; results are presented and discussed in Section 6, while conclusions are given in Section 7, along with the outline of ideas for future work.

2 Related Work

The task of tracking multiple people using video cameras has a long tradition in the field of computer vision; a general overview of the existing multi-target tracking literature can be found in [8]. We focus only on approaches that use *multiple cameras with overlapping fields of view*, as such are usually required for tracking in the world coordinate system. Specifically, we focus on *tracking-by-detection* approaches that rely on robust frame-by-frame detection and are mainly based on the concept of *occupancy maps*. An occupancy map is a plan-view representation of area of interest and allows for efficient aggregation of information coming from different views, usually about the presence of individuals.

The Probabilistic Occupancy Map (POM) by Fleuret *et al.* [5] is a top-down approach; the ground plane is discretized into a probability field in which each cell holds its occupancy probability. A generative model that approximates silhouettes by simple rectangles is used to back-project those probabilities into all views; the occupancy map is obtained by iterative optimization of the probability field, so that the difference between the back-projected and the input binary images is minimized. Berclaz *et al.* [9] use the same framework, but instead of foreground images, the output of a people detector is used. Alahi *et al.* [10] also obtain occupancy map from foreground images in a top-down manner, using sparsity-constrained inverse problem formulation. The work of Khan and Shah [6], on the other hand, is an example of a bottom-up approach. The authors warp the foreground regions from all views into a reference plane, producing a 2-D grid of occupancy likelihoods, which they call a synergy map. Multiple planes parallel to the reference plane are used and the resulting synergy maps are stacked into a 3-D volume representing sampled scene space. A similar approach is also used by Delannay *et al.* [11].

The detections obtained by afore-mentioned approaches can be incorporated into a tracking framework. In [6], graph cuts are used to link the frame-by-frame detections, while in [5], tracking is done using dynamic programming and a local color appearance model. In recent work by Berclaz *et al.* [7], multi-object tracking on top of an occupancy map is formulated as a global optimization problem that can be solved using K-shortest paths algorithm. Their approach completely ignores the appearance and yet it has been shown to outperform state-of-the-art methods. The computational complexity of such approach, however, poses a limit on the amount of frames that can be processed; as reported by [12], the implementation of [7] is, with some modifications, capable of processing 6000 frames, which amounts to 4–5 minutes of video.

While tracking on top of frame-by-frame detections (*tracking-by-detection*) can mitigate the false positive and missing detections that occur in individual frames, it is, from perspective of tracks' identities, *identification-by-tracking*; it propagates identities along the track, and inherently cannot prevent *propagation of identity switches* after people come close together. Therefore, Shitrit *et al.* [13] extended the approach of [7] to preserve consistent identities based on sparse appearance information, namely global color similarity and numbers on the players' jerseys. In this regard, our work is very closely related to [13]. However, whereas

they first perform detection and consider appearance cues in the linking step to ensure consistent identity along the track, we approach the problem from the other side. We perform both detection and identification simultaneously; the obtained detections with identities can then be linked together, resulting in robust *tracking-by-identification*. We deem such an approach more general because it abstracts the integration of additional cues for identifying individuals.

As our work is based on detection and recognition by fusion of multiple weak features, we can find related work in that field as well, especially in single-view object detection and recognition. The most prominent example is work of Breitenstein *et al.* [14], where authors propose the use of general pedestrian detector within a sequential particle filter framework, and combine it with person-specific classifiers that are trained during runtime to distinguish between the targets, based on color and texture features. However, to the best of our knowledge, such fusion is yet to be applied to multi-view, multi-person detection and recognition.

3 A Framework for Feature Map Fusion

A human observer is capable of noticing people even in difficult conditions, by using multiple cues, such as motion, shape and deviation from the background. They can also successfully distinguish between individuals by relying on discriminative visual cues, such as color of clothing, hair color and style, complexion, facial features, body height and width, and even gait and prior knowledge regarding the likelihood of an individual's presence at a certain location. Most of these features are weak, and only their combination allows us to reliably distinguish between individuals. Similarly, our framework uses multiple weak features, encoded as feature maps, and fuses them together using a trained classifier, as shown in Figure 1.

The *feature maps* are our proposed generalization of the well-known concept of an occupancy map [5], where in each cell, instead of probability of occupancy, the value of a feature is stored. The main advantage of using such maps is efficient aggregation of information coming from multiple views and noise reduction due to enforced multi-view constraints. Feature maps allow encoding of different types of features that come from various sources, for example global visual cues of a person's presence, discriminative visual cues for distinguishing between individuals, prior knowledge or information from non-visual sensors (e.g. a radio tracking system).

Note that construction of a feature map does not necessarily require an occupancy map algorithm. For example, if the map is based on prior knowledge, it does not require multi-view constraints and can be constructed directly; same goes for inclusion of information from non-visual sensors. Furthermore, some of the maps are common to all people (e.g. background-subtraction-based detection map), while some (e.g. map of distances to a reference color histogram) are person-specific, meaning that for each person, a separate map is generated.

Since it is expected that usefulness of individual feature maps depends on the situation, the idea is to obtain as many feature maps as possible and fuse them

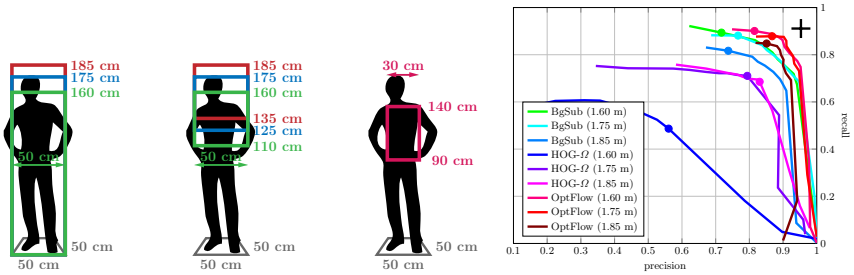


Fig. 2. (a) rectangles for POM based on background subtraction and optical flow, (b) rectangles for POM based on head-and-shoulder (omega shape) detection, (c) sampling region for construction of color histograms, (d) overall precision-recall curves for all global detection cue maps as the σ_{pom} parameter of the POM algorithm is varied. The σ_{pom} parameter controls the desired fitting between the input binary images and back-projected synthetic images; therefore, it influences the compromise between resulting POM's precision and recall. The curves were obtained from training portion of dataset from Section 5.1. The dots on curves denote the operating points we selected for fusion, while the black cross in top-right corner denotes the result of Experiment 1.

using a classifier that has been trained on the annotated training portion of the data. For classification, we stack all feature maps; for each cell, we concatenate the corresponding features into a feature vector and use it as an input to the classifier, in our case a Support Vector Machine (SVM). This way we obtain a SVM score map, on which we perform non-maxima suppression to isolate the person detection(s).

4 Feature Maps

In this section, we describe the feature maps we use to demonstrate the viability of the proposed framework. We outline the reasoning behind using each feature, its characteristics, advantages and disadvantages, and also present some implementation details.

4.1 Global Detection Cue Maps

The global detection cue maps are primarily used to distinguish people from the background. In our case, they are actually occupancy maps produced by publicly-available implementation[1] of the POM algorithm [5], based on different visual cues: background subtraction, optical flow and head-and-shoulder detection.

[1] `http://cvlab.epfl.ch/software/pom/`

Table 1. Per-person recall in all global detection cue maps (in selected operating points), obtained from the training portion of dataset from Section 5.1. For each cue (background subtraction, optical flow, omega shape), three maps are generated using POM algorithm, each assuming a different height of individuals. The recall value depends on the assumed and the actual height of an individual; consequently, the whole stack of global detection maps implicitly encodes some discriminative information about individuals' heights.

Person	Height	Background subtr.			Optical flow			HOG-Omega		
		1.6 m	1.75 m	1.85 m	1.6 m	1.75 m	1.85 m	1.6 m	1.75 m	1.85 m
#1	1.58 m	0.846	0.713	0.594	0.923	0.797	0.748	0.566	0.559	0.301
#2	1.77 m	0.846	0.881	0.839	0.804	0.783	0.790	0.350	0.510	0.524
#3	1.85 m	0.895	0.930	0.874	0.944	0.958	0.930	0.378	0.755	0.846
#4	1.79 m	0.958	0.937	0.881	0.916	0.916	0.853	0.601	0.874	0.867
#5	1.84 m	0.923	0.944	0.895	0.916	0.937	0.916	0.538	0.853	0.888

We use a dense grid, with 0.50×0.50 m cells' centers being placed 0.10 m apart; the resulting overlap allows for more precise localization. For the fixed height of rectangles that model persons' appearance in the POM algorithm, we use three different values (1.6 m, 1.75 m and 1.85 m; see Figure 2a), each yielding a separate map. This way we attempt to account for different heights of individuals; per-person recall values listed in Table 1, obtained from the training portion of dataset described in Section 5.1, show that the global detection cue maps implicitly encode some discriminative information about the heights of individuals.

Figure 2d shows precision and recall curves for all nine resulting maps as the σ_{pom} parameter of the POM algorithm is varied. When selecting an operating point (the σ_{pom} value), we assume that for the fusion of multiple maps, it is preferable to sacrifice some precision for increased recall on individual maps.

It should be noted that the POM algorithm implicitly performs non-maxima suppression, hence its solutions tend to converge to isolated cells. While this is favorable when using a single map, it means that in our case the corresponding detections in different maps can be present in different (although more or less adjacent) cells. To compensate for that, we blur the maps using Gaussian kernel with $\sigma_{blur} = 2.5$ and size of 5 (cells).

Background-subtraction-based POM Feature maps, based on background subtraction, are constructed using the output of algorithm [15], which also comes with shadow detection and removal. However, this comes as a compromise between shadows passing as a part of the foreground and significant fragmentation of the foreground blobs. Based on precision-recall curves (Figure 2d), we use the maps obtained at $\sigma_{pom} = 0.002$. Note that this modality is commonly used in detection step of state-of-the-art tracking-by-detection approaches.

Optical-flow-based POM For the feature maps based on optical flow, we use dense optical flow [16,17] to produce intensity images corresponding to magnitude of displacement, and use those as an input to the POM algorithm. As dense flow is used, we expect the silhouettes to be less fragmented than in case of background subtraction; in addition, the used algorithm models both displacement and illumination changes. On the other hand, it turned out to be somewhat susceptible to moving shadows. Also, as optical flow is computed between consequent frames, the detections are lost as soon as people stop moving. Based on precision-recall curves (Figure 2d), we use the maps obtained at $\sigma_{pom} = 0.002$.

Omega-detection-based POM Another visual cue for person detection is the shape of head and shoulders, the so called omega shape [18], which can be detected from multiple viewing angles. For detection, we use Histograms of Oriented Gradients (HOG) [19] descriptor and a SVM trained on the database provided by [18]. From the detections we generate synthetic images, where detection bounding box is filled by the intensity value corresponding to the probability score given by the SVM; these images are used as input to POM.

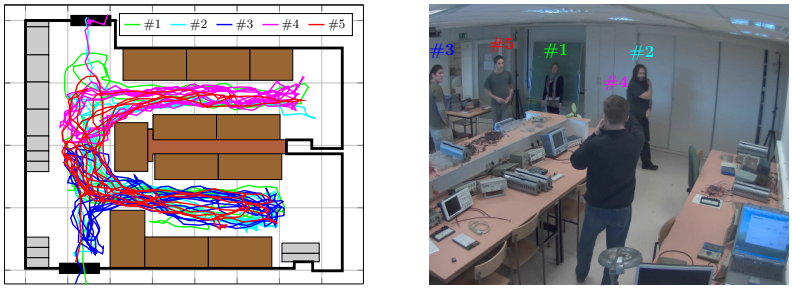
In POM, we restrict the rectangles to the heights at which a face is expected to be found ($1.10-1.60\ m$, $1.25-1.75\ m$ and $1.35-1.85\ m$, respectively, as shown in Figure 2b). As can be seen in Figure 2d, both precision and recall are lower than when using background subtraction images or optical flow. The main reason is that while background subtraction and optical flow can produce fragmented silhouettes, HOG-based detector either detects a head or not, which becomes apparent especially with occlusions. In addition, false positive detections tend to occur on round objects and light reflections on both walls and floor. Based on the precision and recall curves, we use the maps obtained at $\sigma_{pom} = 0.01$.

4.2 Map of Distances to Reference Color Histogram

An often-used discriminative visual cue is color; in our case, we focus on color of the shirts that people wear. To obtain color histograms, we project the corresponding regions ($0.30 \times 0.30\ m$ cells with centers placed $0.10\ m$ apart and height from 0.9 to 1.4 m; see Figure 2c) into views. Using the pixels within projected regions, we construct separate 32-bin histograms for each channel in each view.

From the annotated training data, we obtain average per-view histograms for each person, which serve as a reference. We then construct per-person feature maps by computing Hellinger distance between a person's reference histogram and the histograms of every cell on per-channel and per-view basis; the per-channel distances from different views in which the given cell is visible are then combined using the median operator, resulting in three feature maps. Additionally, we compute the mean distance for all three channels in each view and then combine them using median operator; this way, an additional feature map, representing the average distance to the reference histogram, is obtained.

We use both an RGB-based and an HSV-based maps of distances to a reference color histogram, together amounting to eight feature maps per person.



(a) (b)

Fig. 3. (a) individuals' trajectories, obtained from the training portion of the dataset from Section 5.1, (b) individuals participating in the experiment.

4.3 Location Prior Map

In certain environments, one can also distinguish between (some) people based on the prior knowledge on the likelihood of their presence at a certain location. One example is an office or laboratory where each person has their assigned working place, or certain sports, such as European handball or soccer, where players adhere to positions dictated by their role and the previously-agreed tactics. For example, in European handball, a left-wing player is practically never found in the right wing; such information can help distinguish between multiple players, even though they are visually similar.

We construct the per-person prior maps from the annotated training data (Figure 3a). The annotated positions for a person are placed into a grid and blurred using Gaussian kernel; the resulting map is then rescaled to have its values in range between 0 and 1.

5 Experimental Setup

To the best of our knowledge, there are no publicly available multi-view datasets that would be of sufficient length to verify our proposed approach. In the case of the commonly-used APIDIS dataset [20], a single minute is publicly available. We feel at least another minute from a different part of the sequence would be required for a fair testing. The ISSIA Soccer dataset [21] consists of two minutes of video, but its ground truth is not sufficiently reliable for quantitative evaluation. There are several video sequences offered by the authors of [5], however either they do not come with ground truth, or they lack complete calibration information, which is needed for projecting rectangles at different heights. Furthermore, when ground truth is provided, it is coarse, both spatially and temporally.

Therefore, we present the experimental validation on a challenging dataset that we captured in our laboratory. The dataset presents challenges found both in surveillance scenarios (occlusions both between individuals and by inanimate objects) and in the sports (visual similarity between individuals).

Fig. 4. Views from each of the four cameras.

5.1 The Dataset

We used four cameras to capture a three-minute video sequence at 20 frames per second. The sequence involves five people (Figure 3b) walking around a 7.1×7.0 m room (Figure 4). We manually annotated the ground truth positions on ground plane in each frame. We use every 10th frame of the first minute for training, while all frames from the remaining two minutes are used for testing.

Table 1 shows the variation in individuals' heights. As can be seen in Figure 3b, three individuals wear black and two wear grey clothes, resulting in strong visual similarity between them. Three of the individuals move around the whole room, while the remaining two are confined to either northern or southern part of the room; Figure 3a shows the trajectories obtained from the training portion of the dataset, from which the prior maps are also generated.

5.2 Experiment 1: Improving Generic Person Detection

First, we investigate the benefit of fusing multiple global detection cue maps within our framework only for the purpose of generic person detection and localization. We train a single SVM with nine features (all global detection cue maps); the positive samples are all ground truth points in all training frames, while negative samples are 100 randomly selected non-occupied cells in each training frame. We evaluate the performance of the trained classifier in terms of precision and recall, comparing them to performance of the individual maps.

5.3 Experiment 2: Simultaneous Detection and Identification

In second experiment, we perform simultaneous detection and identification (i.e. detection of specific individual). We independently train and test five SVMs, one for each person, and evaluate their performance. The positive samples are the ground truth points for a particular person in all training frames, while for negative samples cells occupied by other people are combined with 100 other randomly-selected non-occupied cells. The experiment consists of three parts. In the first part (Exp. 2A), only nine global detection cue maps are used. In the second part (Exp. 2B), the maps of distances to reference color histograms are added, and in the third part (Exp. 2C), the prior maps are also used. Such gradual integration of weak discriminative feature maps provides some insight into the extent of their contribution.

5.4 Experiment 3: Tracking-by-Identification

Finally, in order to compare with state of the art, we use the output of our framework to perform tracking-by-identification. In Exp. 3A, we obtain baseline identification-by-tracking results using state-of-the-art KSP algorithm [7] on anonymous detections; because KSP appears to be rather sensitive to large amount of false positive and negative detections in the input data, we use anonymous detections obtained in Exp. 1 instead of raw occupancy maps. In Exp. 3B, we perform tracking-by-identification by separately applying KSP on top of the output of each SVM from Exp. 2C. In both cases, default parameters for KSP are used and no access points are specified.

5.5 Performance Evaluation

For performance evaluation, we use Munkres algorithm to find optimal assignment based on distances between obtained detections and ground truth points; the unassigned detections are considered false positives, while unassigned ground truth points are false negatives. We additionally prevent assignment when distance between a detection and ground truth is greater than 0.5 m to catch gross localization errors. From the obtained statistics on false positive and false negative detections, we compute precision, recall and F-score. When evaluating identification, we also consider whether a detection's and assigned ground truth point's identities match, and compute confusion matrix.

5.6 Classifier

The classifier we use is a CUDA implementation of a Radial Basis Function C-SVM[2], which we modified to return classification scores instead of labels. The optimal values for the SVM parameters C and γ are determined by grid search and five-fold cross-validation on the training data. We assume that the degree to which the provided features are discriminative controls both the resulting precision and recall; in the absence of sufficiently discriminative information, we end up training the SVM with conflicting samples. Therefore, the SVM parameters that result in the highest F-score value are chosen for testing.

6 Results and Discussion

6.1 Experiment 1

With fusion of only global detection cue maps, the overall recall of 91.1% and precision of 95.3% are achieved (black cross on Figure 2d). Comparing to the performance of the individual global detection cue maps, we can see that their fusion outperforms them all both in terms of recall and precision; at comparable recall, the fusion achieves significantly higher precision and vice versa. Therefore our framework can, using multiple cues for detection of a person (even though obtained from the same images), improve generic person detection.

[2] `http://patternsonascreen.net/cuSVM.html`

Table 2. Results of all three parts of Experiment 2. Each SVM is evaluated separately, resulting in a row in the table. In each frame, a SVM can give multiple detections; a detection that is assigned to correct ground truth point contributes to diagonal elements in confusion matrix, whereas detections assigned to ground truth points of other identities contribute to non-diagonal elements. Detections that are left unassigned are considered to be phantoms (see Section 5.5).

		Confusion matrix [%]						Precision	Recall	F-score
		#1	#2	#3	#4	#5	phantom	[%]	[%]	
Exp. 2A	#1	**42.39**	18.30	10.87	15.94	9.06	3.44	42.39	56.80	0.49
	#2	15.71	**19.46**	20.48	20.59	20.14	3.63	19.46	83.25	0.32
	#3	4.95	17.52	**24.82**	24.82	27.25	0.65	24.82	74.27	0.37
	#4	4.97	18.92	23.03	**27.24**	25.84	0.00	27.24	61.17	0.38
	#5	3.88	17.77	24.86	25.61	**27.69**	0.19	27.69	71.12	0.40
Exp. 2B	#1	**83.60**	10.75	0.00	3.76	0.54	1.34	83.60	75.49	0.79
	#2	23.36	**31.52**	3.12	26.59	11.88	3.52	31.52	75.97	0.45
	#3	11.32	9.84	**44.53**	2.46	30.14	1.72	44.53	87.86	0.59
	#4	6.58	26.30	1.23	**50.00**	14.66	1.23	50.00	88.59	0.64
	#5	5.61	15.35	27.77	5.34	**45.13**	0.80	45.13	82.04	0.58
Exp. 2C	#1	**80.05**	9.46	0.51	5.37	0.51	4.09	80.05	75.97	0.78
	#2	27.49	**33.26**	1.28	25.99	8.98	2.99	33.26	75.49	0.46
	#3	4.88	4.66	**62.31**	0.00	25.94	2.22	62.31	68.20	0.65
	#4	10.17	21.66	0.00	**56.25**	5.38	6.54	56.25	93.93	0.70
	#5	7.02	15.79	23.98	7.02	**45.76**	0.44	45.76	75.97	0.57

6.2 Experiment 2

The results of all three parts of the second experiment (the confusion matrix, precision, recall and F-score value) are shown in Table 2. Even using only global detection cue maps (Exp. 2A), some discrimination between the individuals is possible. If it was not, the expected values of precision would be at most 20%; however, for all individuals except #2, the achieved precision exceeds that value. The best precision is achieved for person #1; however, the recall value is relatively low and corresponds to the recall values on the head-and-shoulder detection maps (Table 1) matching this person's height, which indicates that those maps are the main source of discriminative information.

Adding the maps of distances to reference color histograms (Exp. 2B) visibly boosts both precision and recall for all five individuals, because the training samples become more separable due to more discriminative information. If the color-based features were completely reliable (and implicit height information was disregarded), the expected precision values would be 33.3% for individuals wearing black (#1, #2 and #4) and 50% for individuals in grey (#3 and #5). As can be seen, the achieved precision for the individuals wearing black compares favorably to the expected values, whereas for the grey-clad individuals, it is a bit lower. We speculate that this is due to occlusions between people who wear

clothes of different colors, which adversely affects the distances to the reference histograms, especially in cases when a location is covered by only two views. This means that the color feature as used is not robust enough and would benefit either from more views or at least rudimentary occlusion reasoning coming from the POM-based detection maps.

The addition of location prior maps (Exp. 2C) brings some further improvement in precision for the individuals that have informative location priors (#3 and #4; see Figure 3a). For the rest, the addition of a non-informative feature does not significantly affect the results. As can be seen, due to their non-overlapping prior maps, individuals #3 and #4 are not confused by their classifiers anymore. But more importantly, the false positive detections are corrected in cases where other people would be detected by the classifier #3, but they are actually outside the location prior map for person #3.

The overall results of the experiment are very promising; while the obtained classifiers are not yet able to completely distinguish between the individuals, one must bear in mind that the results were obtained on frame-by-frame basis, without any identity propagation whatsoever, and using only a very limited amount of discriminative information. We expect that the results can be significantly improved by integrating additional discriminative cues.

6.3 Experiment 3

Running the KSP algorithm on anonymous detections for baseline state-of-the-art identification-by-tracking (Exp. 3A) results in five trajectories. We assign each trajectory the identity of the person that trajectory is initialized on, and then examine how well that identity is maintained along the trajectory. The algorithm is quite prone to identity switches when people come close, which is reflected in distinctively non-diagonal confusion matrix in Table 3.

In Exp. 3B, separately running KSP on the output of each SVM (tracking-by-identification) results in one trajectory per SVM, together amounting to five trajectories, whose identities are explicitly known. The identities might be switched when people come together, but after they disperse, they are correctly re-established, resulting in much more diagonal confusion matrix (Table 3). The worst results are obtained for person #3, due to many missed detections of their SVM (low recall in Table 2), which cause KSP tracker to drift.

The resulting mean localization error and global identity mismatch rate *gmme* [13] for Exp. 3A and Exp. 3B are 3.17 m vs. 0.53 m, and 0.79 vs. 0.12, respectively. Tracking-by-identification using the output of our framework achieves much better results than state-of-the-art identification-by-tracking, due to ability to properly re-establish identities of individuals.

7 Conclusion

We presented a multi-modal fusion framework for simultaneous person detection, localization and identification from multiple cameras in non-sequential

Table 3. Results of Experiment 3. In each frame, every trajectory produces exactly one point, which is either assigned to correct ground truth point or not; therefore, resulting precision and recall have the same value. Phantoms indicate drifting of the trajectory, which is usually caused by missing detections.

		Confusion matrix [%]						Precision [%]	Recall [%]	F-score
		#1	#2	#3	#4	#5	phantom			
Exp. 3A	#1	**45.78**	0.00	0.00	51.20	0.04	2.98	45.78	45.78	0.46
	#2	6.37	**6.37**	84.20	0.00	0.00	3.05	6.37	6.37	0.06
	#3	18.77	63.11	**7.55**	0.00	6.49	4.08	7.55	7.55	0.08
	#4	5.72	0.84	0.00	**19.53**	72.49	1.41	19.53	19.53	0.20
	#5	18.62	26.06	5.91	25.75	**20.11**	3.55	20.11	20.11	0.20
Exp. 3B	#1	**96.07**	0.11	0.19	0.50	0.50	2.63	96.07	96.07	0.96
	#2	5.23	**76.31**	1.53	10.23	2.33	4.39	76.31	76.31	0.76
	#3	2.06	0.53	**66.27**	0.00	16.67	14.46	66.27	66.27	0.66
	#4	1.34	2.40	0.00	**93.32**	0.00	2.94	93.32	93.32	0.93
	#5	0.84	1.14	0.04	0.11	**95.04**	2.82	95.04	95.04	0.95

manner. Our goal is independent processing of randomly-accessed sections of video, with aim of limiting the unbounded error propagation found in state-of-the-art detection-by-tracking and identification-by-tracking approaches. In our framework, multiple weak features are encoded as feature maps and fused using a trained classifier. Although at this point a limited number of features were used within the proposed framework, we obtained promising results on the task of simultaneous person detection and identification. As demonstrated, the obtained identified detections can be used in tracking-by-identification, which, due to ability to re-establish the identities of individuals, outperformed the state-of-the-art identification-by-tracking approach in terms of average localization error and global identity mismatch rate. Future work will focus on integrating additional discriminative cues within our framework, which we expect to further improve both stand-alone detection and identification, as well as tracking-by-identification results.

Acknowledgement. This work was supported by the research program P2-0095, research project J2-4284 and the research grant 1000-10-310118, all by Slovenian Research Agency.

References

1. Iwase, S., Saito, H.: Parallel tracking of all soccer players by integrating detected positions in multiple view images. In: ICPR 2004. (2004) 751–754 1
2. Xu, M., Orwell, J., Jones, G.: Tracking football players with multiple cameras. In: ICIP 2004. (2004) 2909–2912 1
3. Otsuka, K., Mukawa, N.: Multiview occlusion analysis for tracking densely populated objects based on 2-d visual angles. In: CVPR 2004. (2004) 90–97 1

4. Kristan, M., Perš, J., Perše, M., Kovačič, S.: Closed-world tracking of multiple interacting targets for indoor-sports applications. Computer Vision and Image Understanding **113** (2009) 598–611 1
5. Fleuret, F., Berclaz, J., Lengagne, R., Fua, P.: Multicamera people tracking with a probabilistic occupancy map. IEEE TPAMI **30** (2008) 267–282 2, 3, 4, 5, 8
6. Khan, S., Shah, M.: Tracking multiple occluding people by localizing on multiple scene planes. IEEE TPAMI **31** (2009) 505–519 2, 3
7. Berclaz, J., Fleuret, F., Turetken, E., Fua, P.: Multiple object tracking using k-shortest paths optimization. IEEE TPAMI **33** (2011) 1806–1819 2, 3, 10
8. Yilmaz, A., Javed, O., Shah, M.: Object tracking: A survey. ACM Computing Surveys **38** (2006) 3
9. Berclaz, J., Fleuret, F., Fua, P.: Principled detection-by-classification from multiple views. In: VISAPP 2008. (2008) 375–382 3
10. Alahi, A., Boursier, Y., Jacques, L., Vandergheynst, P.: Sport players detection and tracking with a mixed network of planar and omnidirectional cameras. In: ICDSC 2009. (2009) 1–8 3
11. Delannay, D., Danhier, N., De Vleeschouwer, C.: Detection and recognition of sports (wo)men from multiple views. In: ICDSC 2009. (2009) 1–7 3
12. Ahn, J., Gobron, S., Silvestre, Q., Shitrit, H.B., Raca, M., Pettré, J., Thalmann, D., Fua, P., Boulic, R.: Long term real trajectory reuse through region goal satisfaction. In: Proc. of 4th Intl. Conf. on Motion in Games. (2011) 412–423 3
13. Ben Shitrit, H., Berclaz, J., Fleuret, F., Fua, P.: Tracking multiple people under global appearance constraints. In: ICCV 2011. (2011) 137–144 3, 12
14. Breitenstein, M.D., Reichlin, F., Leibe, B., Koller-Meier, E., Van Gool, L.: Online multiperson tracking-by-detection from a single, uncalibrated camera. IEEE TPAMI **33** (2011) 1820–1833 4
15. Zivkovic, Z., van der Heijden, F.: Efficient adaptive density estimation per image pixel for the task of background subtraction. Pattern Recognition Letters **27** (2006) 773–780 6
16. Werlberger, M., Trobin, W., Pock, T., Wendel, A., Cremers, D., Bischof, H.: Anisotropic Huber-L1 optical flow. In: BMVC 2009. (2009) 7
17. Werlberger, M., Pock, T., Bischof, H.: Motion estimation with non-local total variation regularization. In: CVPR 2010. (2010) 7
18. Li, M., Zhang, Z., Huang, K., Tan, T.: Estimating the number of people in crowded scenes by mid based foreground segmentation and head-shoulder detection. In: ICPR 2008. (2008) 1–4 7
19. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: CVPR 2005. (2005) 886–893 7
20. Vleeschouwer, C.D., Chen, F., Delannay, D., Parisot, C., Chaudy, C., Martrou, E., Cavallaro, A.: Distributed video acquisition and annotation for sport-event summarization. In: NEM Summit 2008: Towards Future Media Internet. (2008) 8
21. D'Orazio, T., Leo, M., Mosca, N., Spagnolo, P., Mazzeo, P.: A semi-automatic system for ground truth generation of soccer video sequences. In: AVSS 2009. (2009) 559–564 8