NOTE!

This paper, titled

Histograms of optical flow for efficient representation of body motion
by Janez Perš, Vildana Sulić, Matej Kristan, Matej Perše, Klemen Polanec and Stanislav Kovačič

was published in the journal *Pattern Recognition Letters*, Volume 31, Issue 11, 1 August 2010, Pages 1369-1376.

The version you have downloaded is in pre-print formatting. For the official version please see the publisher's web site:

http://dx.doi.org/10.1016/j.patrec.2010.03.024

Histograms of Optical Flow for Efficient Representation of Body Motion

Janez Perš *, Vildana Sulić, Matej Kristan, Matej Perše, Klemen Polanec, Stanislav Kovačič

Faculty of Electrical Engineering, University of Ljubljana, Tržaška 25, 1001 Ljubljana, Slovenia

Abstract

A novel method for efficient encoding human body motion, extracted from image sequences is presented. Optical flow field is calculated from sequential images, and the part of the flow field containing a person is subdivided into six segments. For each of the segments, a two dimensional, eight-bin histogram of optical flow is calculated. A symbol is generated, corresponding to the bin with the maximum sample count. Since the optical flow sequences before and after the temporal reference point are processed separately, twelve symbol sequences are obtained from the whole image sequence. Symbol sequences are purged of all symbol repetitions. To establish the similarity between two motion sequences, two sets of symbol sequences are compared. In our case, this is done by the means of normalized Levenshtein distance. Due to use of symbol sequences, the method is extremely storage efficient. It is also performance efficient, as it could be performed in near real-time using the motion vectors from MPEG4 encoded video sequences. The approach has been tested on video sequences of persons entering restricted area using keycard and fingerprint reader. We show that it could be applied both to verification of person identities due to minuscule differences in their motion, and to detection of unusual behavior, such as tailgating.

Key words: Image sequences, Human motion, Optical flow, Levenshtein distance

* Corresponding author.
Tel: ++ 386 1 4768 876
Fax: ++ 386 1 4768 130
Email address: janez.pers@fe.uni-lj.si
(Janez Perš).

1. Introduction

Human motion analysis is important topic in computer vision. In many cases people and their motion form the most informative content of the visual depiction of the scene. This is particularly true for visual surveillance scenarios.

In this paper, we focus on developing the compact representation of human motion, with the primary objective of detecting person-specific behavior when facing access control point, equipped with keycard reader, fingerprint reader and surveillance camera. Our ultimate goal is the ability to verify person's identity using only motion features. Additionally, we aim to detect certain behavior that is not allowed at the control point (entry of multiple persons, also known as *tailgating*, for example).

Most approaches to human identification by motion focused on the problem of recognizing humans by observing human gait (Foster et al. 2003, Wang et al. 2003, Cuntoor et al. 2003, Little and E. Boyd 1998). Human gait is essentially considered as motion of person's legs, while some researchers (Cuntoor et al. 2003) include motion of arms in their gait recognition schemes as well. In 2003, Carlsson (2003) demonstrated that walking people can be recognized from the features, derived by tracking small number of specific points on the human body. He achieved 95% recognition rate on database of 20 recordings of six different persons. Cheng et al. (2008) proposed a method for both automatic path direction and person identification by analyzing the gait silhouette sequence. The gait silhouettes were nonlinearly transformed to low-dimensional embedding and dynamics in time-series images were modelled via HMM in the corresponding embedding space. Laptev et al. (2007) assumed that similar patterns of motion contain similar events with consistent motion across image sequences. They demonstrated that local spatiotemporal image descriptors can be defined to carry important information of space-time events for subsequent recognition.

As shown above, many researchers chose to observe human gait. Human gait is not *any* motion of extremities, it is specifically the motion due to human locomotion (walking, running). The context of locomotion in essence normalizes the observed activity – there are many things people can do with legs and arms, but there are only a few ways a person can walk or run, and the constraints induced by narrowing the context (such as assumption that the gait is periodic) help significantly in the task of gait-based human identification. We rely on the similar effect, which appears in access control scenarios.

Our task of motion-based human recognition is closely related to gesture and activity recognition from images or videos. Published activity recognition algorithms use variety of methods for activity recognition, of which we present only a few. Many more are discussed in detailed surveys, such as Moeslund et al. (2006) and Hu et al. (2004).

Similar to our work, there are several other approaches relying on motion estimation. Black et al. (1997) for example used parametric models on optical flow across the image to estimate facial and limb motion, and recognize facial expressions. Yacoob and Davis (1994) tracked specific regions on the human face and translated them into symbols using a dictionary of universal expressions. Dai et al. (2005) extracted facial action features by observing histograms of optical flow for lower and upper region of the face. Zhu et al. (2006) represented motion in broadcast tennis video by using a new motion descriptor, which is a group of histograms based on optical flow. Motion descriptor based on optical flow measurements in a spatiotemporal volume were used for similarity measure to recognize human actions at lower resolutions by Efros et al. (2003). Laptev et al. (2008) detect points of interest in the spatiotemporal volume and calculate the histograms of gradient and histograms of optical flow in their neighborhood. Histograms are normalized and subsequently concatenated to form feature vectors. Multiple view image sequences are used in Ahmad and Lee (2008), where authors used combination of shape flow and local-global motion flow.

There are approaches to activity recognition that do not rely on tracking or motion estimation. Lu et al. (2008) for example used Histograms of Oriented Gradients (HOG) descriptors to successfully track and recognize the activity of hockey players. The activity recognition was based on the output of HOG descriptor, not the tracking results.

Finally, many authors developed methods, which work on the video data, represented in the form of *spatio-temporal volumes*. Use of motion history volumes (MHV) as a free-viewpoint representation for human actions for example is introduced in Weinland et al. (2006). Their representation can be used to learn and recognize basic human action classes, independently of gender, body size and viewpoint. Mokhber et al. (2008) used global "space–time volumes" composed by the binary silhouettes extracted from each sequence. Actions in their work were therefore represented only by one vector, which permitted usage of simple measurements to determine the similarity between actions and recognize them. Different geometric approach of representation for human actions is described in Yilmaz and Shah (2008), where a set of action descriptors, generated by stacking a sequence of tracked 2D object silhouettes or contours, forms a 3D volume in the spatiotemporal space. Zelnik-Manor and Irani (2006) for example represented image sequences as three dimensional (spatiotemporal) stacks and performed statistical analysis to detect activity boundaries and activity types.

To allow the application of statistical moments to motion based time series analysis, Shutler and Nixon (2006) proposed a new moment descriptor structure that includes spatial and temporal information. They demonstrated the application of the velocity moments using human gait classification, producing a holistic description of temporal motion. Wang and Suter (2008) proposed a general framework to learn and recognize sequential image data in low-dimensional embedding space. To find more compact representations of high dimensional image data, they adopted locality preserving projections (LPP) to achieve the low-dimensional embedding of dynamic silhouette data.

Different from authors mentioned here, Robertson and Reid (2006) had interest in higher-level reasoning about action context in order to develop a system for human behavior recognition in video sequences. They modelled human behavior as a stochastic sequence of actions. Actions were described by a feature vector comprising both trajectory information (position and velocity), and a set of local motion descriptors. Via probabilistic search of image feature databases representing previously seen action, action recognition was achieved.

In this paper, we demonstrate our approach on the task of identifying people by their motion when they approach access control point. Similarly to the gait-based recognition, this task is helped by narrowing down the context of human motion. In our case, people have to perform certain tasks (showing the keycard to the keycard reader and placing a finger on the fingerprint scanner), to gain access. This way, motion is essentially "normalized" to few standard gestures, which provides means for person identification and for detection of unusual behavior. Our approach was designed with practical applications in mind, therefore we placed high importance on the compactness of obtained motion features and the possibility of inexpensive and fast implementation of the proposed method.

2. Our approach

In our preliminary research (Perš et al. 2007), we established that different people behave slightly differently when faced with the need to authenticate themselves to the access control system. Although all persons perform basically the same sequence of tasks (presenting a keycard, placing a finger on a reader, opening of a door), there exist many subtle and less subtle differences in how these tasks are performed. For example, some people carry their cards in the wallets, other in their pockets or purses. Some are left-handed, others are right-handed. Some will come to the access point with the keycard already prepared, others will reach for it in the last moment before authentication. Finally, some will grasp the card with the same hand they use for providing a fingerprint, and others will use both hands. Some people will participate in particular behavior, known as *tailgating*, where one person opens the door, and more persons enter – this is in many cases a violation of access rules and had to be detected as unusual behavior.

To capture those differences between different individuals, and to detect unusual behavior, we developed a method of motion feature extraction, which had to satisfy multiple constraints.

First, to be used in surveillance application, the method of extraction motion features has to be insensitive to lighting, clothing and other circumstances that are beyond our control. This directed the research towards extracting a motion using optical flow, without limiting ourselves to particular implementation of optical flow calculation.

Second, the compact motion representation was needed, for the method to have any chances of ever being used in real world applications, where features of many individuals might be stored in a compact (embedded) device, such as future generations of access point controllers.

Third, the algorithm has to be reasonably fast, to have potential to be used in embedded system without excessive computational power. The computational demands for optical flow calculations are usually high, however, as we will show in the paper, we managed to use MPEG4 compressed streams to obtain motion vectors and therefore bypass the optical flow calculation completely, with good results.

Our approach is based on several assumptions, as follows:

- Cooperative users. We assume that people have vested interest in coming through the access control point with as little hassle as possible. This is not unreasonable, as many other forms of identification require significant cooperation from the user as well (e.g. fingerprint scanners, keycards, iris scanners, just to name a few).
- Existing security policy. We do assume that there are certain rules of behavior that users must adhere to. The task of such a system would be to detect and report the behavior that deviates

from the usual activity.

– Repetitive user behavior. In our preliminary tests, we discovered that after a few weeks of using the access control system, people tend to "optimize" their motion, in a way that is most convenient to them, when faced with an access control point. In an on-line supplement to this paper, at http://vision.fe.uni-lj.si/research/hof/articles/prl09jp/, we present video mosaics of people entering one of the access points as part of their daily routine.

These assumptions allowed us to design a novel method for validating person identity and detecting unusual human behavior at the automated access control points (ACP), based on the descriptors, derived from the histograms of optical flow.

The rest of the paper is structured as follows: first, we will describe the algorithm for comparing video sequences using Histogram of Optical Flow (HOF) descriptors. Then, we will present the system description - the setup in which the test image sequences are captured, along with the HOF implementation details. Following this, we will present the results and conclusions.

3. Histograms of optical flow (HOFs)

Our method is based on extracting motion features from image sequences using optical flow. The distinct advantage of such approach is that the burden of correctly estimating motion in variable lighting conditions and clutter is entirely confined to optical flow calculation. There are many approaches to calculate the optical flow, and as we show in the experimental section, at least two approaches can be used in our framework.

Algorithm 1 summarizes the procedure to obtain HOF motion descriptors from available optical flow field sequences. A frame from one such sequence is shown in Figure 1 a).

This algorithm does not make any assumptions about the source of optical flow data; therefore, it could be applied in variety of ways. The implicit assumption is that the sequences have same frame rate and flow field dimensions. Additionally, the algorithm assumes that each sequence contains a single temporal reference, which can be used for temporal alignment, and that there exists predefined partitioning of the image into sub-regions, such as the partitioning shown in Figure 1 b).

Algorithm 1 : Obtaining HOF descriptors of motion of a person.

Input: Optical flow sequence $F(k)$, definition of n image sub-regions, temporal reference point t_r

Output: HOF descriptor - n sequences of symbols $S_b^i(k)$ and $S_a^i(k)$, $i \in [1, n]$ describing the motion before and after t_r.

Perform temporal smoothing of flow field with the temporal median window spanning each triplet of sequential flow images F_{k-1}, F_k and F_{k+1}, $k \in [2, t_{max} - 1]$

Discard the vectors outside of the predefined region of interest (containing person).

Split the sequence $F(k)$ at the temporal reference point t_r (e.g. key card registration), to the sequences F_b, containing flow before the reference point and F_a, containing the flow after the reference point: $F_b = F(t < t_r)$, $F_a = F(t \geq t_r)$

Initialize $2n$ empty sequences of symbols, n sequences corresponding to the activity before the reference point ($S_b^1 \dots S_b^n$) and n sequences corresponding to the activity after the reference point ($S_a^1 \dots S_a^n$).

for Each flow image $F_b(k)$ and $F_a(k)$, $k \in [2, t_{max}]$ **do**

  Divide the flow field into n sub-regions F^i, as shown in Figure 1 b).

  for Each sub-region i, $i \in [1, n]$ **do**

    Calculate the 2-dimensional histogram $H^i(k, v, \theta) = hist(F^i(k))$ of the optical flow sub-region $F^i(k)$ at the moment k, as illustrated in Figure 1 c). Two histogram dimensions quantize flow amplitude v, and flow direction θ, respectively.

    Find the bin with maximum count in the 2-dimensional histogram, $\underset{v,\theta}{\operatorname{argmax}}(H(k, v, \theta))$

    Generate symbol $s_{v\theta}$, based on a bin with maximum count.

    Add $s_{v\theta}$ to the sub-region symbol sequence, either S_b^i or S_a^i: $S^i \leftarrow \{S^i, s_{v\theta}\}$

  end for

end for

for All sequences S_b^i and S_a^i, $i \in [1, n]$ **do**

  Remove symbol repetitions in the sequence.

end for

The algorithm basically calculates the dominant motion in each of the sub-regions. Both the amplitude and direction of motion are quantized through the use of 2D optical flow histograms, and therefore the dominant motion can be encoded simply by assigning a symbol to each of the histogram bins. This way, a compact representation of whole body motion, including gestures, is built. We call the sets of such symbol sequences *HOF descriptors*. In a real world implementation, the descriptors can be extracted from the flow sequences immediately after the flow is obtained, therefore reducing the need for storage of original video sequences or optical flow field sequences. As described in the next section, this dictionary-based representation of motion can be extremely compact, and is therefore ideally suited for embedded devices.

Observing the maximum in each histogram is inherently noisy approach, however, due to small number of bins, the effects of noise are small. Likewise, the lowest-velocity bin is discarded to get rid of the low-velocity noise, which inevitably appears in optical flow vectors.

In our case, normalized Levenshtein distance in conjunction with nearest-neighbor classification principle is used for sequence comparison. Algorithm 2 summarizes our implementation of HOF descriptor comparison. This approach allows for lightweight implementation of the algorithm, requires no explicit learning, and performs reasonably well, as shown in the Section 5. Levenshtein distance has also been found to be resilient to relatively large amounts of noise (Perše et al. 2009). Other methods could be used as well, provided that certain adaptations are made – most notable alternative are Hidden Markov Models (HMMs).

4. System description and implementation details

Access control points come in many varieties. In our case, the setup consisted of door with electronic lock mechanism, keycard sensor, fingerprint sensor and special controller, connecting all components with the database server. In our setup, we complemented the control point with camera, which observed people entering through the door and recorded image sequences of their pre-entry behavior and the entry itself.

Algorithm 2 : Comparing two HOF motion descriptors in our experiments.

Input: Two sets of HOF descriptor sequences S_α and S_β, each containing $2n$ sequences of symbols, describing the motion in n sub-regions before and after a temporal reference point.

Output: Normalized distance $D(S_\alpha, S_\beta)$ between the descriptors

for All symbol sequences S_b^i, S_a^i, $i \in [1, n]$ **do**

  Calculate the Levenshtein distances L_b^i and L_a^i between the corresponding symbol sequences: $L_b^i = L_b^i(S_{\alpha b}^i, S_{\beta b}^i)$ and $L_a^i = L_a^i(S_{\alpha a}^i, S_{\beta a}^i)$

  Obtain normalized distances Ln_b^i and Ln_a^i by dividing each distance with the length of the longer of the two compared sequences.

end for

Calculate the total normalized distance D between the optical flow sequences as a mean across all distances L_b^i and L_a^i, respectively: $D = \underset{i,x=\{b,a\}}{\text{mean}}(Ln_x^i)$

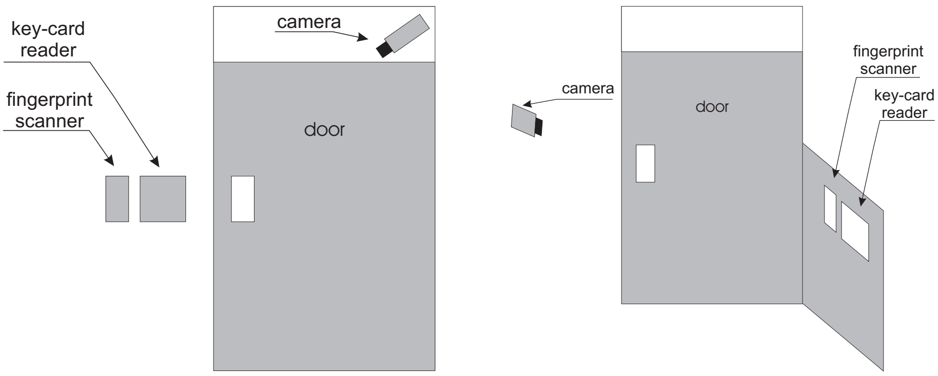
a) b) c)

Fig. 1. a) Optical flow vectors for one frame. b) Scene partitioning, $S_a^i(b)$ denotes sequences $S_a^1 \dots S_a^6$ and $S_b^1 \dots S_b^6$, which are generated by the optical flow in the depicted sub-regions. c) 2-dimensional histograms of optical flow

4.1. *Access control system*

The video acquisition and testing has been done in two locations, and camera and sensor setup differs slightly between the two. In the remainder of the text, we refer to those access control points as *Access control point 1 (ACP 1)* and *Access control point 2 (ACP 2)*, respectively. Fig. 2 shows the positions of sensors and cameras for both locations.

The sequence of activities that each person performed, as they authenticated themselves,



(a) (b)

Fig. 2. Positions of sensors and cameras relative to the door for ACP 1 (a) and ACP 2 (b).

(a) (b)

Fig. 3. Typical view of the person for ACP 1 (a) and ACP 2 (b).

was as follows: approaching the door, presenting the keycard to the keycard reader, waiting for audible confirmation (beep), placing finger on the fingerprint scanner, waiting for audible ”click” of the electronic lock, pulling the door, and finally, entering.

Typical view of the person from each of the cameras is shown in Fig. 3.

4.2. *Data acquisition*

Both test locations were equipped with cameras, with different technology used to capture the videos.

In the case of ACP 1, a 640×480 pixel color IEEE 1394 camera was rotated 90 degrees to better use the available image aspect ratio. Clips, ranging from 8 seconds to 10 seconds at 30 frames per second were recorded using motion detection software to conserve disk space. The recording system was not connected to the access control system, and due to shortcomings of the motion detection scheme, many recordings missed critical elements of the activity and had to be deleted. After a review of videos, 112 complete video clips were selected and manually categorized. Additionally, the videos have been manually temporally aligned with respect to the moment when person's keycard came to closest distance to the keycard reader.

In the case of ACP 2, an Axis 207 indoor video surveillance camera (with resolution set to 640×480 pixels and frame rate to 15 frames per second) was used, and was directly connected to the access control system. Access control system performed *pre-buffering* of the video stream, by storing last 75 video frames in the circular buffer. In the moment when person was successfully authenticated (which corresponds to the moment when person heard a click of the electronic lock), the buffer was stored, and recording continued for further 5 seconds. This way, temporally aligned videos of entries were acquired, and stored on the access control system's database server.

4.3. *Manual data evaluation*

The videos, captured on ACP 1 and ACP 2 were of different nature. At ACP 1, five people were entering the lab as the part of their normal routine. The access control system itself has been in place for six months before camera was installed, and therefore our experiment did not interfere with their activities in any way. Due to consent forms signed by all participants, they were aware that the recordings are taking place.

After the database of videos at ACP 1 has been collected, the videos were visually inspected to evaluate the uniqueness and permanence of motion and activity. We observed that people at ACP 1 indeed developed unique ways of approaching the system, moreover, under the same circumstances they repeatedly performed same sequence of motions to perform authentication.

This rule was broken mainly under the influence of other factors, such as carrying additional objects, tailgating (entry of multiple persons), presence of other people distracting the person who was performing the authentication, and other unusual activities (e.g. leaving the lab door open to return without authentication).

Acknowledging this, videos from ACP 1 were categorized both according to subject identity and subject activity (e.g. person X carrying a bag, person Y carrying a notebook, etc.). Tests at ACP 1 confirmed that people develop unique motion patterns, when they are faced with the task of authentication at the control point.

To further test our approach, ACP 2 was built, and near real-time implementation was tested. At ACP 2, four people were asked to perform complete authentication routine (keycard, fingerprint, entering door) many times. They were asked to perform the required tasks in a way that seemed most convenient for each one of them. While people were entering ACP 1 as part of their daily routine, the tests at ACP 2 were done on several separate occasions, with many entries performed on the same day.

During the online tests at ACP 2, the on-line performance of a presented approach was measured. However, while speed of execution was within our expectations, the classification rate was not. Therefore, videos were archived and inspected. Inspection revealed that the access control system was unable to accurately synchronize many of the video recordings, with delays sometimes exceeding one second. Therefore, videos with improper synchronization were removed from the database and the whole test was re-run in off-line manner using exactly the same algorithm as in the on-line tests.

4.4. *Implementation details*

Tests of the proposed approach were done in two phases.

4.4.1. *Implementation on ACP 1*

First batch of tests was done immediately after all videos from ACP 1 were collected. First, dense optical flow (Black and Anandan 1996) was calculated from image sequences, which were downsampled by the factor of 8 to speed up the calculation. Algorithm 2 was applied as follows. First, median smoothing across temporal axis using three-frame window was applied to reduce noise. Next, optical flow field amplitude was scaled by the factor of 0.48 (0.06 times the downsampling factor), and the 2D histograms of optical flow were calculated for each of the six regions, shown in Fig. 1 b). Histograms for each region were constructed with bin edges of 0, 0.33, 0.66 and 1 in amplitude direction and 0, 90, 180, 270 degrees in angular direction. The contents of the lowest amplitude bins (between 0 and 0.33) were discarded, as they contain mainly noise. Remaining eight histograms were assigned one symbol each, and the sequences $S_b^1 \ldots S_b^6$ and $S_a^1 \ldots S_a^6$ were generated as described in Algorithm 2. Those tests were performed off-line, as dense optical flow calculations require significant amount of computation. Therefore, such approach is accurate, but highly impractical.

4.4.2. *Implementation on ACP 2*

For tests on ACP 2, a near real-time implementation was developed. The system was able to provide distances d_N, as described in Algorithm 2 approximately 15-30 seconds after a person has entered. To achieve this, we used motion vectors, extracted from MPEG4 video stream instead of dense optical flow field. After the recording of each entry was finished, the video was converted to MPEG4 video clip, using widely available open source software encoder (*Mencoder*) and open source *Xvid* codec. Then the clip was immediately decoded using customized version of open source player (*MPlayer*), which extracted motion vector data into separate data file. The process of obtaining motion vectors for all 150 frames took about 10 seconds on 2.4GHz Intel Pentium 4 processor. No subsampling was used, as MPEG4 motion vectors are derived by a block-matching algorithm over a regular grid, and as result, such optical flow is much sparser than dense optical flow by Black and Anandan (1996), which was used in tests at ACP 1.

Such approach allowed us to test the performance of the near real-time prototype implementation. The whole process of extracting HOF descriptors from a sequence and comparing them with a precalculated database of about 100 descriptor sets takes between 15 and 30 seconds. While optical flow was calculated by outside application, the rest of the algorithm was implemented in Matlab and could be significantly optimized, if desired.

4.5. *Experimental setup*

Multiple experiments have been performed on the acquired data. The task of the described prototype system would be to recognize *imposters* (e.g. persons with stolen or borrowed keycard) and, additionally, to detect unusual behavior (e.g. tailgating). To streamline the analysis, HOF descriptors for all videos, acquired on both ACP 1 and ACP 2 were precalculated using the methods described above.

The positions of the cameras at ACP 1 and ACP 2 were significantly different, and the sensor setup (keycard sensor and fingerprint scanner) differed significantly as well. There was also only a slight overlap between the persons entering at ACP 1 and those participating in tests at ACP 2 (one person common to both groups). Therefore, the analysis for ACP 1 and ACP 2 was performed separately.

The videos from ACP 1 have been processed using dense optical flow, while videos from ACP 2 have been processed using MPEG4 motion vectors in place of optical flow. HOF descriptors, obtained as described in the first part of Algorithm 2, were compared to all other descriptors. The normalized distance d_N was observed to assess descriptor performance.

5. Results

5.1. *Results for ACP 1*

Videos of 115 regular entries were classified according to the person and activity (e.g. carrying a bag, carrying a notebook). HOF descriptors from every video have been compared to HOF descriptors of all other videos, and the descriptor with the smallest distance d_N was selected. The results are shown in Table 1. Observing the confusion matrix in Table 1, it can be seen that HOF descriptors identify persons quite well in such setup – in each row, the largest number lies on the matrix diagonal. Success rate (the ratio of properly established identities) was 82% in this case.

Table 1
Confusion matrix for all clips from the database of ACP1. The word after the slash (/) denotes the activity. "plain" denotes the usual mode of authentication - without carrying any objects. "notebook" means that person was carrying a laptop computer, and "bags" means that person was carrying extra luggage. Numbers denote the number of matches between each of the clips in the categories in the first column and categories in the first row.

Person/activity	1/plain	1/notebook	2/plain	3/plain	3/bags	4/plain	5/plain
1/plain	13	0	1	1	0	0	0
1/notebook	2	1	0	1	0	0	0
2/plain	1	0	30	0	0	0	1
3/plain	1	1	1	7	3	0	0
3/bags	0	0	0	1	8	3	0
4/plain	0	0	0	0	0	9	0
5/plain	0	0	2	0	1	1	23

5.2. *Results for ACP 2*

Videos recorded at ACP 2 have been split into four groups. In the first group (Group A), there were videos of 57 regular entries of four test persons. In the second group (Group B) there were videos of 114 regular entries of the same four persons, captured at a later date. In both groups, videos were classified according to identity of a person entering. In the third group (Group C) there were 37 videos of unknown persons. In the last group (Group D) there were 32 videos of tailgatings, which were performed both by known and unknown persons. Recordings from groups C and D were acquired on multiple occasions. In our case, they serve as negative samples. We verified that those groups do not contain any regular entries by the persons participating in videos in groups A and B.

HOF descriptors for all videos have been calculated as described in Sections 4.4.2 and 4.5. For each HOF descriptor from groups A and B the closest match (in terms of the smallest distance d_N) from the same group was found, excluding the comparison to the same descriptor. These results are shown as confusion matrices in Tables 2 and 3. Next, similar analysis was done in cross comparison manner, where for each descriptor, a closest match in the other group was found. These results are shown as confusion matrices in Tables 4 and 5, and show that there is no significant decrease in performance, if videos from one occasion are matched to the videos, acquired at the different occasion. Therefore, we can assume that HOF descriptors, obtained this way are temporally stable to a certain degree.

Again, observing the confusion matrices, it can be seen that HOF descriptors perform well in such setup – numbers on the diagonals are the largest in each row. Success rates for intra-group tests on groups A and B were 91% and 89%, respectively. Success rates for comparison of Group A to Group B and vice-versa were 95% and 85%.

Table 2
Confusion matrix for intra-group test of Group A from ACP2.

Person	1	2	3	4
1	11	0	0	0
2	0	14	0	0
3	0	1	12	2
4	0	1	1	15

All experiments so far were based on pure nearest-neighbor principle. In practice, as number of users would rise, such a system would be faced both with people which are unknown (have no

Table 3
Confusion matrix for intra-group test of Group B from ACP2.

Person	1	2	3	4
1	10	0	1	0
2	0	34	0	3
3	3	1	26	1
4	0	3	0	32

Table 4
Confusion matrix for comparison of Group A to Group B from ACP2.

Person	1	2	3	4
1	11	0	0	0
2	0	14	0	0
3	0	1	13	1
4	0	1	0	16

Table 5
Confusion matrix for comparison of Group B to Group A from ACP2.

Person	1	2	3	4
1	9	0	1	1
2	0	27	3	7
3	1	0	27	3
4	0	1	0	34

existing samples in the database) and people, who perform activities – such as tailgating – that significantly differ from their usual behavior. A practical solution to this problem is addition of threshold-based distance check - checking of a shortest obtained distance against some predefined threshold, and declaring all samples that are above the threshold to be unknown or invalid.

Therefore, in a final experiment, we tested the performance of HOF descriptors in detecting the unknown persons and unusual behavior. For that purpose we compared the minimum distances from the intra-group tests for groups A and B with the minimum distances from groups C and D (unknown persons and tailgatings, respectively) to groups A and B. Figure 4 shows the false negatives and false positives rate, depending on the threshold used. Since the threshold is applied to the distance d_N, lower threshold results in more strict criteria for entry, and higher threshold in more relaxed criteria.

In this context, false negatives denote the cases, where the shortest distance d_N of of a particular entry from groups A or B to the closest (but not the same) entry from groups A or B was higher than a set threshold - in this case, the system would reject a properly behaving person, if it would use groups A and B as the reference for acceptable person's motion. Increasing the threshold naturally lowers the number of such cases. On the other hand, there are two types of false positives: the ones, from group C, where an unknown person would be granted entry, based on shortest distance d_N to any of the entries from groups A or B. These cases are denoted as *false positives "unknown"* in the Figure 4. The other type of false positive occurs, when a system would not detect a tailgating (videos from group D), again, based on the shortest distance d_N to any of the entries from the groups A or B. These cases are denoted as *false positives "tailgating"* in the Figure 4. The number of false positives naturally increases with increasing threshold. Observing Figure 4, it can be seen that the described method is capable of distinguishing between regular and irregular entries. It can be also seen that, if an appropriate threshold is used, for example 0.56, obtained at the intersection of false positives rate for unknown persons and false negatives rate, then the false negatives rate is approx. 20%, false positives rate for unknown persons is approx. 20%, and false positives rate for tailgatings is under 10%.

6. Discussion

We presented Histograms of Optical Flow (HOFs), which were used to compactly describe human motion from sequences. We have shown that HOF descriptors can be used to recognize or verify the identity of the persons in the context of video surveillance, coupled with access control. We have also shown that HOF descriptors can be used to detect unusual and unwanted behaviour, such as entrance of multiple persons using a single keycard - a scenario called "tailgating".

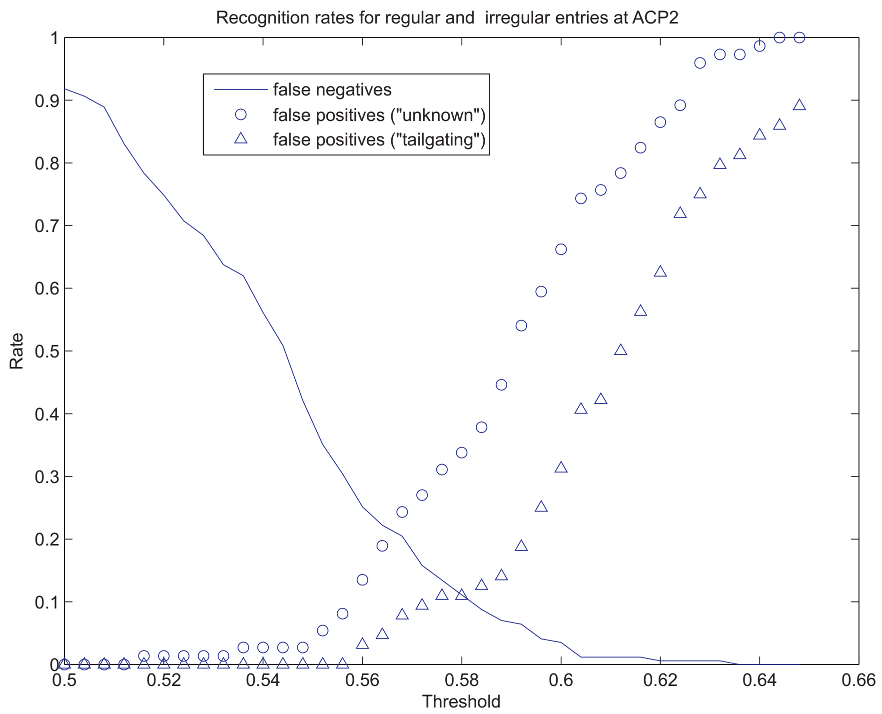


Fig. 4. Recognition rates for regular and irregular entries at ACP 2. There is only one category of false negatives, since it is impossible to determine why certain sample was rejected, other than it was simply too different from the samples from the training set.

The tests have shown that, using currently available off-the-shelf equipment, the results can be obtained in approximately 15-30 seconds, which suffices for near-realtime implementations of our system. The structure of HOF descriptor – a sequence of symbols – allows for very compact representation, which is important for the potential use in embedded devices, such as future generations of access point controllers. With optimized implementation of our method it would perhaps be possible to reduce the overall processing time to the range of few seconds. However, true realtime operation is limited by the fact that a post-authentication part of the video is used for descriptor extraction as well (it does contain motion that is related to person opening the door and entering), and therefore, as presented, cannot be used for realtime decision on whether to grant or deny access to a person currently being identified.

In theory, our method of extracting HOF descriptors is computationally expensive, however most of the computational demands are related to the calculation of optical flow. Currently, there exist algorithms which compute approximations to optical flow, such as motion vectors in MPEG4 compressed sequences, for which we have shown that can be used in our framework. This is important, since there exist hardware MPEG4 compression solutions (such as some network cameras), which would completely eliminate the need for any optical flow computation in our descriptor extraction scheme, provided that the calculated optical flow can be accessed by our algorithm. Since descriptors themselves are extremely compact, and the method of comparing them is simple Levenshtein distance, there is a real possibility of implementing the described scheme in embedded environment.

One drawback of our method is requirement for independent temporal reference. We observe motion that is, in effect, "normalized" (all persons have to perform same task), and the temporal reference (in our case, the moment when keycard is recognized by the access control system) is used to align the sequences, before descriptors are extracted. As we have witnessed in our experiments, even small errors in temporal alignment (e.g. a few frames) can have devastating effect on the recognition rate.

The method can be easily extended to multi-camera setup. Images are divided into segments that are processed separately almost all the way, and only at the end the results are combined in a final distance between sequences. Algorithm itself does not assume any spatial correlation between image segments, therefore, they could as well come from different cameras.

As presented, our method is not well suited to provide *hard decisions* to allow or disallow entry of a certain person. However, such system can decide in near realtime on whether the entry of a person was suspicious or not (either due to wrong identity or other behavioral anomalies), and that information can be used in many ways that are beneficial for the overall security of the protected area. For example, it could be used for alerting the security staff or flagging the log entries for a subsequent or periodic manual security review of the video archive, dramatically improving the efficiency of such undertaking. In that context, it could be used as an automated video database indexing tool.

Although we developed and tested our methodology in the framework of an access control point scenario, we believe that the method has potential for a wider use, especially in situations where people are expected to perform certain tasks, and the deviation from their tasks is sufficient reason for alarm. Most of those scenarios involve people interacting with machines in one or another way, which also provides opportunity for obtaining above mentioned temporal reference. Two of the examples are:

– People operating heavy machinery – for example a person interacting with forge press, where the sequence of operations is clearly defined, however, people are often tempted to take dangerous shortcuts.
– People interacting with high-tech equipment, where in interest of safety, certain procedures have to be followed. A example of this are pre-flight and pre-landing checklists on a flight deck of a passenger airplane; pilots are required to perform certain sequence of predefined activities, and many of these activities include motion. Absence of such activities in any case hints to a dangerous situation on a flight deck. Conversely, unexpected activity during other phases of the flight may be sufficient reason for a silent alarm as well.

Acknowledgement

The research, presented in this paper has been supported by Slovenian Ministry of Defence (MORS) contracts CIVaBis (M2-0156), PDR (M3-0233 C), and, in part by Slovenian Research Agency (ARRS) contracts P2-0095 and P2-0232.

References

Ahmad, M. and Lee, S. W.: 2008, Human action recognition using shape and clg-motion flow from multi-view image sequences, *Pattern Recognition* (41), 2237–2252.

Black, M. J. and Anandan, P.: 1996, The robust estimation of multiple motions: Parametric and piecewise-smooth flow fields, *Computer Vision and Image Understanding* **63**(1), 75–104.

Black, M. J., Yacoob, Y. and X. Ju, S.: 1997, Recognizing human motion using parameterized models of optical flow, *in* M. Shah and R. Jain (eds), *Motion-Based Recognition*, Kluwer Academic Publishers, Boston, pp. 245–269.

Carlsson, S.: 2003, Recognizing walking people., *I. J. Robotic Res.* **22**(6), 359–370.

Cheng, M.-H., Ho, M.-F. and Huang, C.-L.: 2008, Gait analysis for human identification through manifold learning and hmm, *Pattern Recognition* (41), 2541–2553.

Cuntoor, N., Kale, A. and Chellappa, R.: 2003, Combining multiple evidences for gait recognition, *Multimedia and Expo, 2003. Proceedings of ICME '03*, pp. III: 113–116.

Dai, Y., Shibata, Y. and Cai, D.: 2005, Understanding facial expressions by the hierarchical recognition of genuine emotions, *International Journal of Innovative Computing, Information and Control* **1**(2), 203–214.

Efros, A. A., Berg, A. C., Mori, G. and Malik, J.: 2003, Recognizing action at a distance, *IEEE International Conference on Computer Vision–ICCV'03*, Nice, France, pp. 726–733.

Foster, J. P., Nixon, M. S. and Prugel-Bennett, A.: 2003, Automatic gait recognition using area-based metrics, *Pattern Recogn. Lett.* **24**(14), 2489–2497.

Hu, W., Tan, T., Wang, L. and Maybank, S.: 2004, A survey on visual surveillance of object motion and behaviors, *IEEE Transaction on Systems, Man, and Cybernetics–Part C: Applications and reviews* **34**(3), 334–352.

Laptev, I., Caputo, A., Schueldt, C. and Lindeberg, T.: 2007, Local velocity- adapted motion events for spatio-temporal recognition, *Computer Vision and Image Understanding* (108), 207–229.

Laptev, I., Marszalek, M., Schmid, C. and Rozenfeld, B.: 2008, Learning realistic human actions from movies, *Proc. Int. Conf. Computer Vision and Pattern Recog. (CVPR'08)*, Anchorage, Alaska, pp. 1–8.

Little, J. J. and E. Boyd, J.: 1998, Recognizing people by their gait: The shape of motion, *Videre: Journal of Computer Vision Research* **1**(2).

Lu, W.-L., Okuma, K. and Little, J. J.: 2008, Tracking and recognizing actions of multiple hockey players using the boosted particle filter, *Image and Vision Computing – In press* .

Moeslund, T., Hilton, A. and Krűger, V.: 2006, A survey of advances in vision-based human motion capture and analysis, *Computer Vision and Image Understanding* (104), 90–126.

Mokhber, A., Achard, C. and Milgram, M.: 2008, Recognition of human behavior by space-time silhouette characterization, *Pattern Recognition Letters* (29), 81–89.

Perše, M., Kristan, M., Kovačič, S., Vučkovič, G. and Perš, J.: 2009, A trajectory-based analysis of coordinated team activity in a basketball game, *Computer Vision and Image Understanding* **113**(5), 612–621.

Perš, J., Kristan, M., Perše, M. and Kovačič, S.: 2007, Motion based human identification using histograms of optical flow, *in* M. Grabner and H. Grabner (eds), *Computer Vision Winter Workshop 2007*, St. Lambrecht, Austria, pp. 19–26.

Robertson, N. and Reid, I.: 2006, A general method for human activity recognition in video, *Computer Vision and Image Understanding* (104), 232–248.

Shutler, J. D. and Nixon, M. S.: 2006, Zernike velocity moments for sequence-based description of moving features, *Image and Vision Computing* (24), 343–356.

Wang, L. and Suter, D.: 2008, Visual learning and recognition of sequential data manifolds with applications to human movement analysis, *Computer Vision and Image Understanding*

(110), 153–172.

Wang, L., Tan, T., Ning, H. and Hu, W.: 2003, Silhouette analysis-based gait recognition for human identification, *IEEE Trans. Pattern Anal. Mach. Intell.* **25**(12), 1505–1518.

Weinland, D., Ronfard, R. and E., B.: 2006, Free viewpoint action recognition using motion history volumes, *Computer Vision and Image Understanding* (104), 249–257.

Yacoob, Y. and Davis, L.: 1994, Computing spatio-temporal representations of human faces, *Proceedings CVPR '94*, Seattle, WA, USA, pp. 70–75.

Yilmaz, A. and Shah, M.: 2008, A differential geometric approach to representing the human actions, *Computer Vision and Image Understanding* (109), 335–351.

Zelnik-Manor, L. and Irani, M.: 2006, Statistical analysis of dynamic actions., *IEEE Trans. Pattern Anal. Mach. Intell.* **28**(9), 1530–1535.

Zhu, G., Xu, C., Huang, Q. and Gao, W.: 2006, Action recognition in broadcast tennis video, *18th International Conference on Pattern Recognition–ICPR'06*, Hong Kong, pp. 251–254.