Selective Covariance-based Human Localization, Classification and Tracking in Video Streams from Multiple Cameras

A. R. Taranyan, V. V. Devyatkov and A. N. Alfimtsev
Bauman Moscow State Technical University, Moscow, Russian Federation

Keywords: Pattern Recognition, Computer Vision, Human Tracking, Covariance Matrix, Covariance Region Descriptor, Selective Localization.

Abstract: In this paper a novel selective covariance-based method for human localization, classification and tracking in video streams from multiple cameras is proposed. Such methods are crucial for security and surveillance systems, smart environments and robots. The method is called selective covariance-based because before classifying the object using covariance descriptors (in this case the classes are the different people being tracked) we extract (selection) specific regions, which are definitive for the class of objects we deal with (people). In our case, the region being extracted is the human head and shoulders. In the paper new feature functions for covariance region descriptors are developed and compared to basic feature functions, and a mask, filtering out the most of the background information from region of interest, is proposed and evaluated. The use of the proposed feature functions and mask significantly improved the human classification performance (from 75% when using basic feature functions to 94.6% accuracy with the proposed method) while keeping computational complexity moderate.

1 INTRODUCTION

Reliable real-time tracking of moving objects using multiple cameras, wherein each camera has relatively small field of view, is a challenging task that becomes particularly difficult to solve when traced object speed increases or such object temporarily disappears. It has a wide range of applications, including security and surveillance systems (Watada, 2008), smart environments, robotics (Bellotto, 2009) and human-computer interaction (Devyatkov, 2011). Before taking on task definition of this paper and its comparative analysis against existing researches, we are going to introduce a number of terms.

An image formed by a digital video camera at a given time t will be called a frame. A frame captured at the moment t will be denoted as I_t. A video stream will stand for a frame sequence I_t, I_{t+1}, ... , I_{t+k}. Let us denote a single pixel of a frame I_t as $I_t(x, y)$. Entries $I_{Rt}(x, y)$, $I_{Gt}(x, y)$ and $I_{Bt}(x, y)$ will denote R, G and B components of the pixel $I_t(x, y)$ accordingly.

A subset of pixels of a frame I_t, containing an object to be identified (for example a person's face), will be called the region of interest (ROI). The region of interest of a frame I_t will be denoted as R_t. A set of pixels of a frame, which does not belong to the region of interest will be called background.

Localization of an object is the process of determining the location of the region of interest containing this object in the frame at a given time t.

Tracking of the object is a process of sequential localization of $R_t, R_{t+1}, \ldots, R_{t+k}$ ROIs containing the object being tracked in $I_t, I_{t+1}, \ldots, I_{t+k}$ frames.

Classification of ROI is an implicit partitioning of a set of all these regions into subsets called classes (In our case, the classes are the people being tracked). Classification requires a class descriptor, which at first is calculated based on certain features of the chosen object that belongs to the class; and later the descriptor is used for classification as a model to which a descriptor calculated under the same features and procedure (a sample) is compared. A variety of methods has been dedicated to the classification of ROIs, which mainly differ from each other by the types of the descriptors used. All these methods are usually divided into two large groups.

The methods of the first group, when forming the descriptor, use key point extraction, extraction of

Taranyan, A., Devyatkov, V. and Alfimtsev, A.
Selective Covariance-based Human Localization, Classification and Tracking in Video Streams from Multiple Cameras.
DOI: 10.5220/0006538100810088
In *Proceedings of the 11th International Joint Conference on Biomedical Engineering Systems and Technologies (BIOSTEC 2018) - Volume 3: BIOINFORMATICS*, pages 81-88
ISBN: 978-989-758-280-6

specific features of the object that describe its geometrical singularities in the best possible way (angles, arcs, lines, contour shapes etc.). Among the methods of this group, we can distinguish the use of SIFT descriptors (Lowe, 2004; Fazli, 2009) that are based on extraction of turn and scale invariant characteristics; the "shapecontext" method (Belongie, 2002); finding of typical parts (particularly human body parts) (Ioffe, 2001), silhouettes (Elzein, 2003) etc. A substantial drawback of these methods is the need of high resolution of the objects being classified, while the method developed in this paper can function effectively even with video streams having comparatively low resolution.

The descriptors of the second group, to classify the ROI, use low-level features (gradients, colours, intensities, positions, etc.) calculated over the whole ROI. The most common descriptors of this group are histograms (Liu, 2014; Comaniciu, 2003) and covariance matrices (Tuzel, 2006; Wu, 2009a; Hassen, 2015).

A disadvantage of histograms as descriptors, at least in known researches, is the dropping of spatial information, which makes impossible to distinguish objects of similar colours but with different shape. The covariance matrix, unlike histograms, generally can be constructed for any number of both colour and spatial features, while keeping calculation complexity moderate. This descriptor is quite popular nowadays, and has a wide range of applications. For example, Ergezer (2016) used the covariance matrix as a descriptor of person's path to solve a problem of anomalous human behaviour detection, and Sanin (2013) applied it for action and gesture recognition. Owing to the mentioned advantages, in this paper the covariance matrix is chosen as a descriptor for the developed method of human localization, classification and tracking by several cameras.

It should be noted that our method, without any preliminary training, should be able to track the same person from multiple cameras and in multiple locations, under various angles, with various backgrounds. As opposed to object tracking from a single camera, there is a significant limitation for application of the methods based on the assumption that the object being tracked has nearby coordinates in the subsequent frames of the video stream (Elzein, 2003). While, in case of tracking from a single camera, existing methods show quite good results, in our case application of such methods is possible only for each particular video stream separately.

2 ALGORITHM OF COVARIANCE-BASED TRACKING FROM MULTIPLE CAMERAS

The proposed method is based on the application of the Viola-Jones classifier for localization of all regions of the peoples' head and shoulders and further matching (classification) of the localized regions to the people that had been detected before via the covariance descriptor. The head and shoulders region has been chosen as a rather informative for people classification, but, meanwhile, a region, that is rarely occluded by other objects.

The main algorithm's pseudocode is presented in Fig. 1. At each moment of time we read frames from all the cameras, localize head and shoulders regions on that frames, then construct covariance matrices for the localized regions. Every localized head and shoulders region can either be a new person, who hasn't been detected yet, or a new occurrence of previously detected person. Thus, we compare that person's covariance descriptor with the descriptors of all of the previously detected people (D) and find the closest one (d*). If it is close enough to the new descriptor, we register a new occurrence of a person, who corresponds to the descriptor d*. Otherwise, we register a new person detection, and add its covariance descriptor to the set D.

In the proposed method, two modules can be distinguished: the module of localization of the head and shoulders regions, and the module of classification (hereinafter referred to as the classifier). Essentially, the classifier is the complex of the chosen method of descriptor construction and the descriptor matching method.

We shall consider the key stages of the proposed algorithm of covariance-based tracking from multiple cameras and discuss in detail the proposed classifier.

2.1 Localization of the Head and Shoulders Regions

Localization of the head and shoulders regions (ROIs) at a given time t is carried out for each video stream separately; then, the detected regions are joint in a common set of regions $R_1, R_2, \ldots, R_m$, which is the very result of this stage. Localization of the ROIs in each particular video stream is performed in two stages. At the first stage static regions of the frame are being discarded by means of

```
Initialize detected people's descriptors D = ∅
FOR t = 1 to n
      R = ∅
      Read I_t^1, I_t^2, …, I_t^p frames from all cameras
      FOR j = 1 to p
            Localize R_j1, R_j2, …, R_jk
            R = R ∪ {R_j1, R_j2, …, R_jk}
      FOR EACH region r ∈ R
            Construct covariance descriptor d for r
            IF D == ∅
                  Register new person detection
                  D = D ∪ {d}
            ELSE
                  d* = Closest(d, D)
                  IF Dist(d*, d) < threshold
                        Register detection of the person, that has descriptor d*
                  ELSE
                        Register new person detection
                        D = D ∪ {d}
```

Figure 1: The pseudocode of the algorithm of covariance-based tracking from multiple cameras.

the background subtraction algorithm based on the Gaussian distribution mixture (Zivkovic, 2004); after that the head and shoulders regions are localized at the remaining frame regions using the Viola-Jones algorithm (Viola, 2004). The Viola-Jones algorithm has been chosen due to its high performance and good precision; thereby it is one of the most widely used algorithms when considering object localization tasks.

Application of the background subtraction allows to increase the localization quality due to the reduction of the number of false-positive detections in the static frame regions (where there are no people or people have already been localized when entering the frame), as well as, in some scenarios, increasing the localization speed by scanning only the part of the frame by the classifier.

2.2 Classifier. Construction of Covariance Descriptors

In this paper the covariance matrix (CM) described by Tuzel (2006) is proposed to be used as a ROI descriptor. The covariance matrix, as previously noted, allows fusing the interconnections of different features of the ROI, among which the main ones are certainly the colour and spatial information. The diagonal entries of the covariance matrix represent the variance of each feature and the non-diagonal entries represent the correlations.

The CM for the region of interest R is constructed in the following way:

1. For each pixel $p_i \in R$ a feature vector $f_i = F(p_i)$ of dimension d is calculated via predefined feature function $F: R \to R^d$, that contains the information on the pixel and, perhaps, on certain region around the current pixel. The function may look like as follows:

$$F(x,y) = [x\ y\ I_R(x,y)\ I_G(x,y)\ I_B(x,y)\ \partial_x(x,y)\ \partial_y(x,y)] \quad (1)$$

   Where x, y are the coordinates of the pixel, $I_R(x,y)$, $I_G(x,y)$, $I_B(x,y)$ are the values of the components R, G and B of the pixel with coordinates (x, y), ∂_x and ∂_y are the first derivatives of the intensity at the point (x, y) in horizontal and vertical directions accordingly.
2. For the resulting set F_R of feature vectors of the region R a mean vector f_{mean} is calculated, and the covariance matrix of dimension d x d is constructed:

$$C_R = \frac{1}{n-1}\sum_{i=1}^{n}(f_i - f_{mean})(f_i - f_{mean})^T, \quad (2)$$

   where n is the number of pixels in the region R.

The complexity of the covariance matrix construction is $O(nd^2)$, where n is the number of pixels in the image (in our case 60×55), and d is the number of components in the selected feature function.

Another important advantage of using the covariance matrices as descriptors is that they are low-dimensional compared to other descriptors. Covariance matrix has only $(d^2 + d)/2$ different values, whereas if we represent the same region via joint feature histogram we will need m * d elements, where m is the number of histogram bins for each feature.

Fig. 2 demonstrates the covariance matrices constructed for two different images using the feature function (1).

It is obvious that the key aspect in the process of construction of a covariance matrix is the selection of the feature function F. The section 2.4 of this paper is dedicated to the problem of selection of the function F.

The proposed descriptor can be enhanced by analyzing the significance of specific pixels of the region of interest when constructing the covariance matrix. Since the rectangular region that is detected by the Viola-Jones algorithm when finding the head and shoulders of a person, contains quite large regions essentially being a background, it makes sense to construct a covariance matrix only for the pixels belonging to the head and shoulders region of the person.

For this purpose, a binary mask M with the size of 60×55 pixels has been calculated semi-automatically on the training set by means of the developed greedy algorithm; the zeroes in this mask (black pixels) denote pixels which are not representative when constructing a descriptor, and the figures of one (white pixels) are the pixels which should be used when constructing a covariance matrix (fig. 3).

Thus, we suggest constructing the covariance matrix C_R of the region of interest R based on feature vectors only of the pixels that correspond to the figure of one in the mask M.

Figure 3: The mask of an image.

2.3 Classifier. Descriptor Matching

To determine the "similarity" of descriptors of two ROIs, the metric for covariance matrices has to be determined. In this paper the efficiency of using of the Euclidian measure D_{eucl} and the measure D_{eigen}, based on the generalized eigenvalues for the covariance matrices being compared (Tuzel, 2006), has been analysed.

D_{eucl} for the covariance matrixes C_1 and C_2 is calculated in the following way:

$$D_{eucl}(C_1, C_2) = \sqrt{\sum_{i=1}^{d}\sum_{j=1}^{d}(C_1(j,i) - C_2(j,i))^2} \quad (3)$$

Calculation of the metric D_{eucl} has the complexity $O(d^2)$.

The metric D_{eigen} is calculated in the following way:

$$D_{eigen}(C_1, C_2) = \sqrt{\sum_{i=1}^{d} ln^2 \lambda_i(C_1, C_2)} \quad (4)$$

where $\{\lambda_i(C_1, C_2)\}_{i=1...d}$ are non-zero generalized eigenvalues for the matrices C_1 and C_2, calculated for the equation

$$\lambda_i C_1 x_i - C_2 x_i = 0,\ i = 1 \ldots d \quad (5)$$

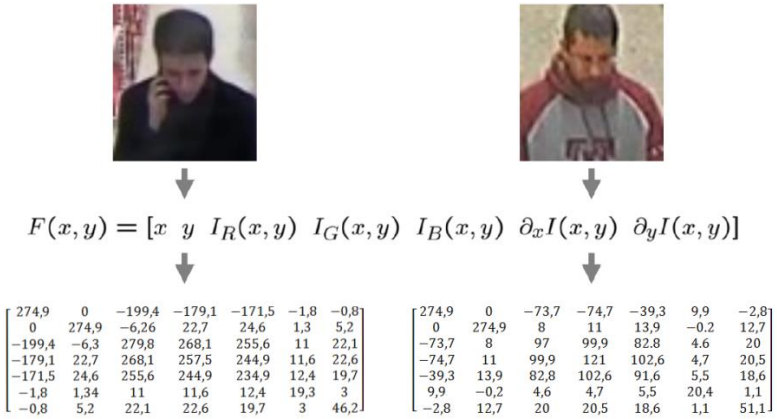

Figure 2: Covariance matrices for two images.

It should be noted that non-negativity of the generalized eigenvalues $\{\lambda_i(C_1, C_2)\}_{i=1...d}$ follows from the fact of covariance matrices being positive-semidefinite matrices.

Calculation of the metric D_{eigen} has a complexity of $O(d^3)$, conditioned by the calculation of generalized eigenvalues.

2.4 Classifier. Selection of the Feature Function

Since the representativeness of a covariance matrix is directly determined by the selection of a feature function F(x, y), a detailed comparison of the efficiency of using of previously proposed feature functions has been carried out, as well as new feature functions have been developed and tested.

As a matter of convenience, the elements of the vector being defined by the feature function F(x, y) have been divided in two subsets:

$$\begin{aligned} &F(x,y) \\ &= [a_1(x,y) \dots a_t(x,y)\ b_1(x,y) \dots\ b_p(x,y)] \\ &= [A(x,y)\ B(x,y)] \end{aligned} \quad (6)$$

where $a_i(x, y)(i = 1 \dots t)$ represent the information on the colour, and $b_i(x, y)(i = 1 \dots p)$ represent the spatial information, with $t + p = d$.

For the colour component A two trivial schemes have been considered that represent the information on the colour of the pixel in RGB and HSV models:

$$\begin{aligned} &A_{RGB}(x,y) = \\ &[x\ y\ I_R(x,y)\ I_G(x,y)\ I_B(x,y)] \end{aligned} \quad (7)$$

$$\begin{aligned} &A_{HSV}(x,y) = \\ &[I_H(x,y)\ I_S(x,y)\ I_V(x,y)] \end{aligned} \quad (8)$$

However, testing of these schemes showed that they don't perform good enough, and to increase the informativeness of the data on the colour features of the ROI being encoded, a scheme $A_{RGBHistN}$ based on histograms has been developed. To construct the feature vectors of this scheme, the ROI is divided into a grid with 5 rows and 5 columns. Then, for each cell of this grid, for each of the components R, G and B, fuzzy histograms are constructed with N bins, after which the feature vectors themselves are constructed:

$$\begin{aligned} &A_{RGBHistN}(x,y) = \\ &[I_R(x,y) I_G(x,y) I_B(x,y) \\ &HR_1(x,y) \dots HR_N(x,y) \\ &HG_1(x,y) \dots HG_N(x,y) \\ &HB_1(x,y) \dots HB_N(x,y)] \end{aligned} \quad (9)$$

where $HR_i(x, y), HG_i(x, y)$ и $HB_i(x, y)$ are the values of i-th bins of the R, G and B histograms of the grid cell, that contains the pixel (x, y).

The idea behind the suggested $A_{RGBHistN}$ scheme is to encode much more detailed colour information on the ROI, than the basic A_{RGB} and A_{HSV} schemes do. Furthermore, this scheme allows encoding the correlation between the cell of the grid and its colour.

Note that the feature vectors for pixels, belonging to the same cell, share the same histogram data, while having their own $I_R(x, y)$, $I_G(x, y)$, $I_B(x, y)$ values.

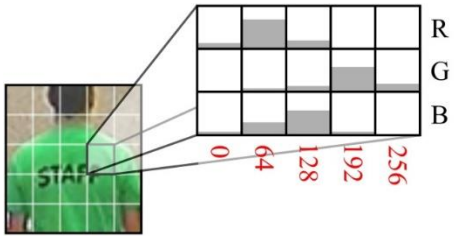

Figure 4: The histograms for one of the cells of the ROI for the feature vector $A_{RGBHist5}$.

Fig. 4 demonstrates the histograms for one of the cells of the ROI (representing the head and the shoulders of a person wearing green outdoor clothes) constructed for the feature vector $A_{RGBHist5}$.

The B_{Radial}, B_{Coord}, B_{Deriv} schemes have been considered as a spatial element B of the feature function F, and a new scheme $B_{RadialGrid}$ has been proposed.

The value of the vector B_{Radial} for the pixel represents the distance of this pixel from the center of the ROI:

$$B_{Radial}(x,y) = [\sqrt{(x - x_0)^2 + (y - y_0)^2}] \quad (10)$$

where x_0 and y_0 are the coordinates x and y of the center of ROI.

The scheme B_{Coord} encodes the information on the coordinates x and y:

$$B_{Coord}(x,y) = [x\ y] \quad (11)$$

The scheme B_{Deriv}, in addition to the coordinates x and y, also encodes the information on the first derivatives of intensity of the pixels of the ROI at the point (x, y) along the axes x and y:

$$B_{Deriv}(x,y)\ [x\ y\ \partial_x(x,y)\ \partial_y(x,y)] \quad (12)$$

The developed scheme $B_{RadialGrid}$ is intended to encode the correlation between the ROI colour information on the ROI area.

For this purpose, the ROI is divided into a grid with 5 rows and 5 columns, and the $B_{RadialGrid}$ scheme is defined in the following way:

$$B_{RadialGrid}(x,y) = [\sqrt{(x-x_0)^2+(y-y_0)^2}\ R_1(x,y)\ R_2(x,y)\ R_3(x,y)\ R_4(x,y)\ R_5(x,y)\ C_1(x,y)\ C_2(x,y)\ C_3(x,y)\ C_4(x,y)\ C_5(x,y)] \quad (13)$$

where $R_i(x,y) = 1$ if the pixel (x, y) belongs to the i-th row, and $R_i(x,y) = 0$ otherwise; and, in the same manner, $C_j(x,y) = 1$ if the pixel (x, y) belongs to the j-th column, and $C_j(x,y) = 0$ if it does not.

Thus, combining the reviewed colour and spatial schemes, 16 feature functions have been formed for the covariance descriptor of the ROI.

As previously noted, to construct a classifier, a descriptor for the ROI and a descriptor matching method have to be chosen. In this paper the construction of a covariance descriptor has been described in detail, a mask has been proposed, and 16 various feature functions for the covariance descriptor have been developed.

Besides, 2 different metrics for covariance descriptors have been reviewed. By combining feature functions, application or non-application of the mask, as well as the metric used, we will have 64 various classifiers for classification of the head and shoulders regions. The following section contains the experimental comparison of these classifiers, based on which the most efficient classifier for the current problem has been chosen.

3 EXPERIMENTAL RESULTS

To test the developed method of human classification based on the images of the head and the shoulders, a test dataset has been formed containing 413 images of the head and the shoulders of 93 different people (hereinafter referred to as 93 classes) (Taranyan, 2017). This dataset was created on the basis of video tapings obtained by the authors, the PETS 2006 dataset of people images, as well as images from the Internet. The reason we've extended the PETS dataset was to make it more challenging – for example, we have added people images, whose clothes have similar colours, but different patterns. Fig. 5 demonstrates test images of 4 different people.

Figure 5: Examples of 4 different classes of the dataset.

The elements of all 93 classes were compared to each other inside the classes, thereafter all possible pairs of classes were considered, and comparison of the elements among the classes of these pairs was carried out.

The quality of each classifier was measured by selecting a value of the threshold, for which the balanced classification rate (BCR) of that classifier is maximal. BCR is defined as the average of true positive rate and true negative rate. This maximal BCR of the classifier is hereinafter referred to as the quality of the classifier.

At first, we have compared the performance of the classifier with the feature function $A_{RGB}B_{Radial}$ using D_{eigen} metric, using the proposed mask, and without using it. The results showed a significant increase in classification quality when using the proposed mask – the quality of the classifier with the mask was 81.8%, and the quality of the classifier without the mask was 75.0%. All the following tests were carried out using the proposed mask.

Then, we have compared the efficiency of using the metrics D_{eucl} and D_{eigen}. The metric D_{eucl}, while being the faster one, also showed a better classification quality (87.5% compared to the 81.8% shown by D_{eigen}). Thus, all the following tests were carried out using the metric D_{eucl}.

Fig. 6 shows the results of comparison of the 16 classifiers based on feature functions that represent all the possible combinations of four colour schemes A and four spatial schemes B. On the figure the classifiers are grouped by the colour scheme A.

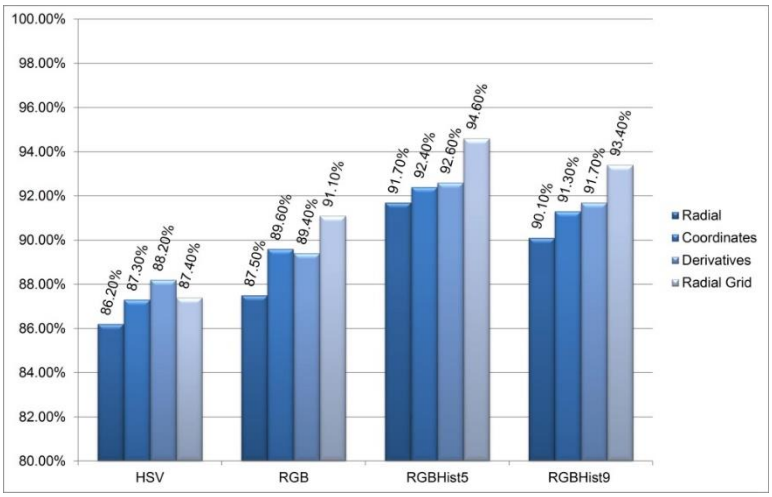

Figure 6: Comparison of quality of the classification for the feature functions being reviewed.

As we can see, the overall performance of the feature functions based on $A_{RGBHist5}$ and $A_{RGBHist9}$ colour schemes is significantly better, then the performance of the feature functions based on A_{HSV} and A_{RGB}.

Among the feature functions based on $A_{RGBHist5}$ and $A_{RGBHist9}$, the best result was achieved using the proposed feature function $A_{RGBHist5}B_{RadialGrid}$, that showed the classification quality of 94.6%. It should be noted that the results showed that it is unreasonable to increase the dimension over 5 in the $A_{RGBHistN}$ scheme.

The results of the feature function comparison have shown, that the proposed feature function allowed creating a strong classifier, that not only can distinguish persons wearing clothes of different colours, but also differentiates persons wearing clothes of similar colours but having different patterns.

Table 1: Performance speed of classifiers for feature functions being reviewed (milliseconds per comparison).

Classifier	Performance speed
HsvDerivatives	1.5 ms/comp
RgbRadialGrid	1.4 ms/comp
RgbHist5RadialGrid	4.1 ms/comp
RgbHist9RadialGrid	8 ms/comp

In the Table 1 the comparison of the performance speeds of the classifiers is presented (benchmarked on Intel Core i5-6500 running at 3.2GHz). The classifier based on $A_{RGBHist5}B_{RadialGrid}$, although not being the fastest one, still shows a decent speed, allowing to perform up to 250 ROI pair comparisons per second, where each comparison includes covariance matrices construction for each ROI of the pair and the comparison of the covariance matrices.

Based on the classification quality and the performance speed, we think that the classifier, based on the $A_{RGBHist5}B_{RadialGrid}$ scheme and Euclidean metric, is optimal for human head and shoulders region classification. The complexity of this classifier is $O(nd^2)$ for covariance matrix construction and $O(d^3)$ for covariance matrix comparison, where n is the number of pixels in ROI, and m is the number of features in the $A_{RGBHist5}B_{RadialGrid}$ feature function.

4 CONCLUSIONS

In the paper the task of human tracking, localization and classification in video streams from multiple cameras has been reviewed. Solution of this task is crucial for development of video surveillance and security systems, smart environments, robots and systems with human-computer interaction. We have proposed a method of human localization, classification and tracking in video streams from multiple cameras, which incorporates a selective mask and is based on covariance descriptors. The proposed method increased the human classification efficiency from 75% to 94.6%, which is a quite good result taking into account the complexity of the used dataset. The key feature of the proposed method is the possibility to classify people based on the covariance descriptor omitting the training stage.

We have proposed and evaluated two novel ideas for feature function selection for covariance matrices:

- Splitting the ROI into grid, construction of histograms for grid cells and assignment of the

same histogram information to feature vectors of the pixels from the same cell.

- Encoding in covariance matrix the correlation between ROI cell colour information and the position of the cell in the grid.

In the process of development of the method presented in this paper, we tested a mask allowing getting rid of a large part of the pixels of the ROI which are background, selected the metric for covariance descriptors that is most appropriate for this task, reviewed common feature functions, developed new ones and carried out a detailed experimental analysis of their efficiency.

The method can be further improved by incorporating a frame-to-frame prediction (particle filter, for example) for each particular video stream separately, and by using adaptive descriptors, which encode information on multiple occurrences of the person, and are being updated during the person tracking.

ACKNOWLEDGEMENTS

This work was supported by the Ministry of Education and Science of the Russian Federation R & D State project №2.5048.2017/8.9.

REFERENCES

Bellotto, N., & Hu, H. (2009). Multisensor-based human detection and tracking for mobile service robots. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics), 39(1), 167-181.*

Belongie, S., Malik, J., & Puzicha, J. (2002). Shape matching and object recognition using shape contexts. *IEEE transactions on pattern analysis and machine intelligence, 24(4), 509-522.*

Comaniciu, D., Ramesh, V., & Meer, P. (2003). Kernel-based object tracking. *IEEE Transactions on pattern analysis and machine intelligence, 25(5), 564-577.*

Devyatkov, V., & Alfimtsev, A. (2011). Human-computer interaction in games using computer vision techniques. *In Business, Technological, and Social Dimensions of Computer Games: Multidisciplinary Developments (pp. 146-167).* IGI Global.

Elzein, H., Lakshmanan, S., & Watta, P. (2003, June). A motion and shape-based pedestrian detection algorithm. *In Intelligent Vehicles Symposium, 2003. Proceedings. IEEE (pp. 500-504). IEEE.*

Ergezer, H., & Leblebicioğlu, K. (2016, October). Anomaly Detection and Activity Perception Using Covariance Descriptor for Trajectories. *In European Conference on Computer Vision (pp. 728-742).* Springer International Publishing.

Fazli, S., Pour, H. M., & Bouzari, H. (2009, December). Particle filter based object tracking with sift and color feature. In Machine Vision, 2009. *ICMV'09. Second International Conference on (pp. 89-93). IEEE.*

Hassen, Y. H., Ouni, T., Ayedi, W., & Jallouli, M. (2015, January). Mono-camera person tracking based on template matching and covariance descriptor. In Computer Vision and Image Analysis *Applications (ICCVIA), 2015 International Conference on (pp. 1-4). IEEE.*

Ioffe, S., & Forsyth, D. A. (2001). Probabilistic methods for finding people. *International Journal of Computer Vision, 43(1), 45-68.*

Liu, H., Wang, L., & Sun, F. (2014). Mean-Shift Tracking Using Fuzzy Coding Histogram. *International Journal of Fuzzy Systems, 16(4).*

Lowe, D. G. (2004). Distinctive image features from scale-invariant keypoints. *International journal of computer vision, 60(2), 91-110.*

Sanin, A., Sanderson, C., Harandi, M. T., & Lovell, B. C. (2013, January). Spatio-temporal covariance descriptors for action and gesture recognition. *In Applications of Computer Vision (WACV), 2013 IEEE Workshop on (pp. 103-110). IEEE.*

Taranyan, A. (2017). Human Tracking Dataset. Available at: https://github.com/Taranyan/HumanTracking-*DataSet* [Accessed 20 May 2017].

Tuzel, O., Porikli, F., & Meer, P. (2006). Region covariance: A fast descriptor for detection and classification. *Computer Vision–ECCV 2006, 589-600.*

Viola, P., & Jones, M. J. (2004). Robust real-time face detection. International journal of computer vision, 57(2), 137-154.

Watada, J., & Musaand, Z. B. (2008, August). Tracking human motions for security system. *In SICE Annual Conference, 2008 (pp. 3344-3349). IEEE.*

Wu, Y., Cheng, J., Wang, J., & Lu, H. (2009, September). Real-time visual tracking via incremental covariance tensor learning. In Computer Vision, 2009 *IEEE 12th International Conference on (pp. 1631-1638). IEEE.*

Zivkovic, Z. (2004, August). Improved adaptive Gaussian mixture model for background subtraction. In Pattern Recognition, 2004. ICPR 2004. *Proceedings of the 17th International Conference on (Vol. 2, pp. 28-31). IEEE.*