HUMAN DETECTION IN GROUPS USING A FAST MEAN SHIFT PROCEDURE

Csaba Beleznai[1]*, Bernhard Frühstück*[2] *and Horst Bischof* [3]

[1]Advanced Computer Vision GmbH - ACV, Vienna, Austria
csaba.beleznai@acv.ac.at
[2]Siemens AG Österreich, Programm- und Systementwicklung, Graz, Austria
[3]Institute for Computer Graphics and Vision, Univ. of Technology, Graz, Austria

ABSTRACT

Detecting individual humans within groups becomes a non-trivial task when performing automatic visual surveillance in crowded scenes. This paper proposes a novel way to detect individual humans directly from the difference image using a fast variant of the mean shift mode seeking procedure and verifying the hypothesized configuration by a model-based approach. The method runs in real-time. Promising results are demonstrated for challenging image sequences.

1. INTRODUCTION

Detection of humans is an extensively investigated topic in the field of automated visual surveillance. In crowded scenes, a large number of interacting humans usually form groups where individuals become partially or completely occluded. In such cases, detecting individual humans becomes a challenging task. Common approaches of "blob" detection [1, 2] based on thresholding and subsequent morphological filtering of the difference image segment groups as single objects due to the underlying connected component labelling (see Figure 1.b and Figure 1.c). Approaches using segmentation based on color only [3] are not likely to yield good results since often colors are not representative for different individuals. Approaches to detect individuals in groups have been proposed using silhouette analysis [2, 4] and stochastic segmentation from binary images [5]. These approaches require good motion segmentation quality in order to robustly find certain landmark points such as heads along the silhouettes. Methods exploring the solution space of possible human configurations [5] are also computationally expensive, thus unsuitable for real-time applications.

In this paper, we propose a novel real-time method to detect humans in a crowded scene based on the mean shift procedure. Our method does not rely on thresholding for motion segmentation, but it directly operates on the difference image. Similarly to the method proposed by [6], the difference image is interpreted as a two-dimensional probability distribution function where high intensities imply high probabilities for moving objects. Detecting humans thus becomes a clustering problem. Unlike [6] explaining the difference image by mixtures of Gaussian clusters, we use the more general mode-seeking property of the mean shift procedure [7] as a robust nonparametric clustering technique suitable to detect significant modes (see Figure 1.d).

This work has been carried out within the K plus Competence Center ADVANCED COMPUTER VISION. This work was funded from the K plus Program.

(a) (b) (c) (d)

Fig. 1. (a) Sample input frame, (b) motion segmentation using a common approach, (c) regions found by a common approach, (d) humans found by the proposed method.

The paper is structured as follows: In Section 2, we describe how the mean shift procedure is used for mode seeking within the multimodal difference image. In Section 3, we introduce a novel fast version of the mean shift procedure relying on integral images and capable to detect a large number of modes in real-time. Section 4 describes a model-based approach to validate detected modes. Section 5 presents and discusses the experimental results. Finally, the paper is brought to conclusion in Section 6.

2. CANDIDATE LOCALIZATION BY MEAN SHIFT

The mean shift analysis is a relatively recent clustering approach, which can delineate regions belonging to high density modes within a complex feature space. The method was originally proposed by Fukunaga and Hostetler [8]. Cheng [9] generalized the method and pointed out, that the mean shift algorithm is a mode-seeking process on the density function surface. Comaniciu and Meer [7] proved the convergence of the iterated mean shift procedure on discrete data, proposed several extensions and presented its benefits for practical applications.

The mean shift vector for discrete data can be represented in general by the difference between the weighted mean computed with kernel profile $g(x)$ and $\mathbf{x}$ (the kernel or window center):

$$\mathbf{m}(\mathbf{x}) = \frac{\sum_{i=1}^{n} \mathbf{x}_i g\left(\left\|\frac{\mathbf{x}-\mathbf{x}_i}{h}\right\|^2\right)}{\sum_{i=1}^{n} g\left(\left\|\frac{\mathbf{x}-\mathbf{x}_i}{h}\right\|^2\right)} - \mathbf{x} \ , \qquad (1)$$

where n is the number of data points and h is the size of the kernel. See [7] for more details on mean shift clustering.

The difference image is the outcome of a change detection process by forming the difference between the current image of an im-

age sequence and a reference image representing the background. The difference image generated for a crowded scene usually contains large number of high-intensity peaks or modes. Our principal objective is to find individual modes representing possible human candidates. The search process is facilitated by information of expected scaling of humans $\{H(y_i), W(y_i)\}$ at a given vertical image location y_i, which can be obtained by a rough calibration of the scene [10]. H and W are the height and the width of a human, in pixels.

Mode detection within the difference image I is performed by the following steps:

1. The difference image intensity maximum is mapped to unit intensity and its entire range is scaled proportionally.

2. A sample set of n points $X_1 \dots X_n$ is defined by locating local maxima. Local maxima are found by: (1) locating the global intensity maximum and adding it to the list of sample set; (2) resetting the difference image intensity around the found maximum within a window of size $(0.5W(y), 0.5H(y))$; (3) repeating the maximum search of step (1) until the found maximum drops below a threshold T_1.

The points of the sample set are subsequently used in a mean shift procedure. The final result does not depend critically on T_1. A very low value just increases the run time and generates more outliers which have to be eliminated by the model-based mode validation step.

3. The mean shift procedure is applied to the points of the sample set with a window size of $(W(y), H(y))$ according to the local scaling. For the two-dimensional probability distribution in the difference image, the mean shift vector (m_x, m_y) computation using uniform kernel can be defined as:

$$m_x(x,y) = \frac{\sum_{i=1}^{n} x_i I(x_i, y_i)}{\sum_{i=1}^{n} I(x_i, y_i)} - x \tag{2}$$

$$m_y(x,y) = \frac{\sum_{i=1}^{n} y_i I(x_i, y_i)}{\sum_{i=1}^{n} I(x_i, y_i)} - y \tag{3}$$

Starting out from the points of the sample set, the mean shift vector is computed repeatedly until convergence, locating the closest mode typically within 3-4 iterations. Note that the mean shift vector computation requires the computation of the zeroth and first moments of the distribution within the window. The computed zeroth moment can be interpreted as the probability density sampled at distinct locations (x, y) of the convergence path:

$$p(x,y) = \frac{1}{W(y)H(y)} \sum_{i=1}^{n} I(x_i, y_i) \tag{4}$$

The above formula yields a relative measure for the presence of a human-sized moving region. The set of probability measures $\{p(x_1, y_1), \dots, p(x_k, y_k)\}$ computed for the points of the convergence path is used later on to validate individual human locations as it will be explained in Section 4.

The size of the sample set is usually on the order of several hundred points, which represents considerable computational complexity when targeting real-time operation. To achieve real-time mode seeking we introduce a fast variant of mean shift computation using integral images. The technique of fast mean shift computation is described in Section 3.

4. The convergence points of individual mean shift procedures are linked together forming the centers of detected clusters. Linking is carried out analogously to [7] by merging all points which are closer in $x-$ and $y-$direction than the local window size $(W(y), H(y))$. Cluster center coordinates are computed as the weighted mean of convergence points $\{\mathbf{cx}_1, \dots, \mathbf{cx}_r\}$ using the weights $\{p(\mathbf{cx}_1), \dots, p(\mathbf{cx}_r)\}$. The convergence trajectories leading to the cluster center define the paths of the cluster. A bounding box around the cluster paths gives a simple representation for the basin of attraction delineating the region covered by the cluster. Subsequently it is checked whether an overlap between basins of attraction is present. In the case of an overlap, overlapping basins are merged and thus a cluster might possess multiple cluster centers. For a group of people in the difference image, the basin of attraction covers the entire group and - for well-separated modes - individual cluster centers indicate candidate locations for single humans in the group.

3. FAST MEAN SHIFT COMPUTATION

The integral image - also called summed area table [11] - provides an efficient means to compute area integrals for a given image. The integral image at a location (x, y) contains the cumulative sum of pixels located to the left and above (x, y) including the pixel (x, y):

$$I_{int}(x,y) = \sum_{x' \le x, y' \le y} I(x', y') \ , \tag{5}$$

where $I(x, y)$ is an arbitrary image and $I_{int}(x, y)$ is the integral image, i.e., a the two-dimensional cumulative distribution function of $I(x, y)$. Fast computation of the integral image can be performed in a single pass [12]. Using the integral image, the area sum of a rectangular region within the original image can be efficiently computed by the following step:

$$\begin{aligned} S_{area} = {} & I_{int}(x-1, y-1) + I_{int}(x+W-1, y+H-1) \\ & - I_{int}(x-1, y+H-1) - I_{int}(x+W-1, y-1), \end{aligned} \tag{6}$$

where S_{area} is the area sum within the rectangle with upper left corner (x, y) with (w, h) parameters for width and height.

Efficient computation of the mean shift vector (see equations 2 and 3) can be achieved by computing integral images of the first moments:

$$I^x_{int}(x,y) = \sum_{x' \le x, y' \le y} x' I(x', y') \ , \tag{7}$$

$$I^y_{int}(x,y) = \sum_{x' \le x, y' \le y} y' I(x', y') \ . \tag{8}$$

The above quantities can be computed in the same pass as the zeroth moment (Eq. 5). Using the integral images of the zeroth and the first moments the mean shift vector components applying a uniform kernel can be expressed as:

$$m_x(x,y) = \frac{S_{area}(I^x_{int})}{S_{area}(I_{int})} - x \ , \tag{9}$$

$$m_y(x,y) = \frac{S_{area}(I^y_{int})}{S_{area}(I_{int})} - y \ . \tag{10}$$

The numerator and denominator of the above expressions denote the area sums computed over the window rectangle $(W(y), H(y))$ using the precomputed integral images for zeroth (I_{int}) and first (I^x_{int}, I^y_{int}) moments. Given that a single sum computation based on integral image requires three arithmetic operations and four array accesses, the fast computation of mean shift vector $\mathbf{m}(x, y)$

takes nine arithmetic operations and twelve array accesses only. In addition, the computational complexity is independent of the window size. For small window sizes, there is no significant speedup when compared to straightforward summation-based mean shift computation. For larger window sizes (tested up to 90-by-90 pixels), our comparisons yielded a speedup factor of up to 30.

4. MODEL-BASED VALIDATION

The mode detection using mean shift provides stable human candidate positions for single isolated modes of the difference image intensity distribution. However, partially occluded humans often generate contaminated intensity distributions where only the prevailing mode is detected (see Figure 2). Validation of a hypothesized human configuration can be used to best explain the observed distribution within each delineated cluster, i.e. the basin of attraction. Validation is performed using a simple human model to find the configuration of the maximum likelihood. The human model consists of three rectangular regions as shown in Figure 2.

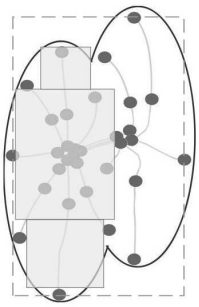

Fig. 2. Illustration of the model-based validation process. The intensity distribution is represented by the touching elliptical regions. Mean shift convergence paths are drawn as grey lines, points represent positions during mean shift convergence. The basin of attraction is shown as a dashed rectangle.

We use a penalized likelihood criterion to select the models best explaining the data. The basic idea is to incrementally generate models for each cluster. Subsequently, the likelihood of the models is calculated using the number of trajectory points explained by the model.

A given arrangement of Z models yields the configuration θ_Z. To efficiently determine the optimum configuration θ_Z^{opt}, we exploit the fact that the mean shift convergence process slows down when close to a local minimum of the density gradient. Thus, local plateaus in the difference intensity (i.e. partially occluded humans) are marked by the accumulation of convergence path points. Therefore, instead of searching the entire cluster region R_C, search is constrained to discrete points of the convergence path $\{(x_1, y_1), \ldots, (x_k, y_k)\}$ with computed local probability densities $\{p(x_1, y_1), \ldots, p(x_k, y_k)\}$ (see Eq. 4). A configuration likelihood based on the intensity distribution within the cluster is computed to compare alternative configurations:

$$P(I|\theta) = \exp\Bigg(-a \left[1 - \frac{1}{A_{R_\theta}} \sum_{x,y \in R_\theta} I(x,y) \right] - b \left[\frac{1}{A_{R_U}} \sum_{x,y \in R_U} I(x,y) \right] \Bigg), \quad (11)$$

where R_θ is the region explained by models. $R_U = R_C \setminus R_\theta$, denoting the region within the cluster not represented by models. A_{R_θ} and A_{R_U} are the areas of the regions R_θ and R_U, respectively. Equation 11 formulates a likelihood measure inversely proportional to the amount of missing measurement (first term) and the unexplained observation (second term). Scaling parameters a and b were estimated by experiments.

The likelihood function of Eq. 11 is used in an incremental model insertion procedure:

(1) A model at the detected mode is inserted;

(2) Cost of the actual configuration is computed

$$C(\theta_z) = (1 - P(I|\theta_z))e^{\beta Z} \ . \quad (12)$$

The last term in the above expression penalizes configurations with large number of models Z. An adequate scaling term β was determined in experiments.

(3) An additional model is inserted at the location where the highest local probability density assigned to convergence path points can be found, considering only path points which are still not explained by a model. The cost of the configuration θ_{Z+1} with $Z+1$ inserted models is computed.

(4) If the configuration with an additional model is favored, the model is kept. The procedure is repeated from step 2 until there is no drop in the configuration cost. The resulting configuration is kept as optimum configuration best explaining the intensity distribution within the cluster.

5. EXPERIMENTAL RESULTS AND DISCUSSION

Background differencing was carried out using a standard method [1] and the obtained difference images were used for mean shift-based detection. The proposed approach was tested on numerous data sets. The presented results are obtained for each frame independently without using any tracking information.

Results are presented for four image sequences depicting crowded scenes and exhibiting frequent occlusions between people. Typical frames of the sequences are shown in Figure 3 and in Figure 4.

Sequence A consists of 7730 frames with a resolution of 360-by-288 pixels (Figure 3, left image). Detected humans are illustrated by superimposed rectangular human models. Each model has a confidence rating. Models with low configuration likelihood (Eq.12) are shown in dark gray, as it can be seen at the top region of the frame, where people are entering or leaving the scene, thus they are only partially visible. Occasional false positives are generated by moving objects other than humans such as a rotating newspaper stand.

Sequence B contains 370 frames with a resolution of 320-by-240 pixels (Figure 3, right image). It depicts a scene where the height of humans in the image ranges from couple of pixels (close to the horizon) to about 25 pixels. Detection was performed within the region where expected height of humans was more than five pixels. Stable results are obtained for humans with a height larger than ten pixels, which shows the excellent mode sensitivity of the proposed method even for poorly resolved regions.

Detection results for sequences C and D consisting of 1500 frames (360-by-288 pixels) are shown in Figure 4.

Detection performance was evaluated using the sequences A, B, and C. Ground truth data was generated by manually annotating a set of frames for each sequence (see Table 1). Moving humans

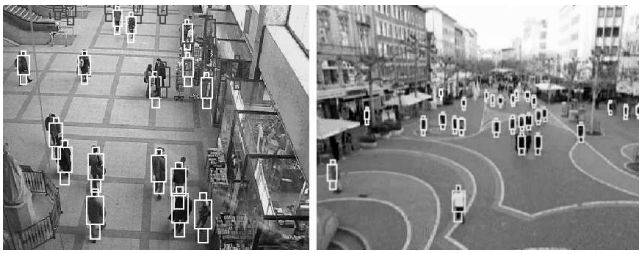

Fig. 3. Human detection results in the sequences A and B.

Fig. 4. Human detection results in the sequences C and D.

with more than 50% visible parts were considered as valid. Correct detection was assumed when ground truth position was inside of the central (torso) bounding box of detected human. Low confidence detections were ignored during evaluation. A one-to-one mapping between ground truth data and detection results was enforced.

Sequence	A	B	C
No. of annotated frames	470	187	879
Valid humans	6147	4749	5380
Detection rate	87.9%	81.2%	84.3%
False alarm rate	29%	12.1%	19%

Table 1. Detection performance of the proposed method given for a range of frames in sequences A, B and C.

Detection results in terms of detection rates and false positive rates are shown in Table1. Detection rates exceeding 80% are obtained for all test sequences. Some missed detections occur when the human's clothing has colors similar to the background. Occasional missed detections also occur, when the saddle point between distributions - representing two or more partially overlapping humans - is found as a mode; an error which is currently not remedied by the model validation step. The relatively high false positive rates stem from the fact, that the difference image often contains motion clutter in form of slightly swaying stationary humans, reflections or shadows leading to occasional false positives. False positives are detected sporadically, thus the integration of detection results into a tracking framework would eliminate the majority of false positives.

The fast detection technique was implemented in C++ and tests were carried out on a 2.5 GHz PC. We obtained frame rates between 12 and 15 fps, depending on the number of detected humans in a frame. Significant portion of processing power is consumed by the motion detection algorithm.

The presented approach uses only few parameters which do not influence the final outcome to a great extent. Surprisingly, even a very simple human model - enabling fast hypothesis-and-test cycles - provides an effective means for human detection in groups.

6. CONCLUSIONS AND FUTURE WORK

We have presented a novel approach to detect humans in a crowded scene based on a fast variant of the mean shift method in real time. Stable results are produced and even poorly resolved (≈ 10 pixels) groups are well explained by simple human models.

Data-driven model selection will be needed to render the method sufficiently generic for scenes containing also objects other than humans. Incorporating the detection procedure into a tracking framework is also subject to future work.

7. REFERENCES

[1] R. Collins, A. Lipton, T. Kanade, H. Fujiyoshi, D. Duggins, Y. Tsin, D. Tolliver, N. Enomoto, and O. Hasegawa, "A system for video surveillance and monitoring: VSAM final report," in *Technical Report CMU-RI-TR-00-12*, Robotics Institute, Carnegie Mellon University, 2000.

[2] I. Haritaoglu, D. Harwood, and L. S. Davis, "W4: Real-time surveillance of people and their activities," *IEEE Trans. on PAMI*, vol. 22, no. 8, pp. 809–830, 2000.

[3] A. Elgammal, R. Duraiswami, and L. S. Davis, "Efficient nonparametric adaptive color modeling using fast gauss transform," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, December 2001, vol. 2, pp. 563–570.

[4] Y. Kuno, T. Watanabe, Y. Shimosakoda, and S. Nakagawa, "Automated detection of human for visual surveillance system," in *Int. Conf. on Pattern Recognition*, Vienna, Austria, August 1996, p. C92.2.

[5] T. Zhao and R. Nevatia, "Bayesian human segmentation in crowded situations," in *IEEE Conference on Computer Vision and Pattern Recognition*, Madison, USA, June 2003, pp. 459–466.

[6] A.E.C. Pece, "Tracking by cluster analysis of image differences," in *Proceedings of the 8th Int. Symposium on Intelligent Robotic Systems*, Reading, UK, July 2000.

[7] D. Comaniciu and P. Meer, "Mean shift: A robust approach toward feature space analysis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 5, pp. 603–619, 2002.

[8] K. Fukunaga and L. D. Hostetler, "The estimation of the gradient of a density function, with applications in pattern recognition," *IEEE Transactions on Information Theory*, vol. 21, pp. 32–40, 1975.

[9] Y. Cheng, "Mean shift, mode seeking, and clustering," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 17, no. 8, pp. 790–799, 1995.

[10] G. A. Jones, J. R. Renno, and P. Remagnino, "Auto-calibration in multiple-camera surveillance environments," in *Third IEEE International Workshop on Performance Evaluation of Tracking and Surveillance*, Copenhagen, Denmark, June 2002, pp. 40–47.

[11] F. Crow, "Summed-area tables for texture mapping," in *Proceedings of SIGGRAPH*, 1984, vol. 18, pp. 207–212.

[12] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, Kauai, Hawai, 2001, vol. 1, pp. 511–518.