PLAYER EXTRACTION IN SPORTS VIDEO SEQUENCES

Slawomir Mackowiak, Jacek Konieczny

Chair of Multimedia Telecommunications and Microelectronics,
Poznań University of Technology
email: jkonieczny@doctorate.put.poznan.pl, smack@et.put.poznan.pl

ABSTRACT

This paper presents the results of the work on developing methods for object segmentation and preparation of algorithms for tracking objects in 3D space. This paper focuses on the case where video sequences recorded by many, fast-moving cameras with changing focal length lenses are processed. It describes in detail two main processes of the segmentation algorithm, namely the global camera motion estimation and object model fitting, using calculation of camera parameters from homography matrix. In the paper, the results for player detection in sport video sequences are presented.

Index Terms—segmentation, video surveillance, stereoscopic video, sport video sequences

1. INTRODUCTION

Although detection and tracking of objects is commonly known in literature, most of the existing approaches assume specific conditions such as fixed cameras, single moving object, and relatively static background. In sports video broadcasts, such strict conditions are not applicable. Firstly, the cameras that are used to capture sports games are not static and they are in almost permanent motion. A broadcasted video is the one selected according to the broadcast director's instruction from frequent switches among multiple cameras. Thirdly, there are numerous players moving in various directions in the broadcasted video. Finally, the background in sports video changes rapidly. Those conditions make detection and tracking of objects in broadcasted video difficult.

The main goal of the paper is to present the algorithm dedicated to sports application where many cameras, many different shots, many different lightning conditions and fast moving objects exists in a sequence together.

According to the applications and the different characteristics (complexity, accuracy of segmentation of still and moving areas, rapid motion of objects, shadow, nonuniform lighting, required operator assistance and occlusion), the analysis of segmentation methods for moving objects in a video sequences is done. Below is the list of most common techniques for segmentation of fast moving objects that could constitute a basis of the segmentation process: adaptive Gaussian mixed models (GMM) [4], methods based on gathering information about the energy transmitted in the content of the sequence [5], morphological tools [6-7], background modeling [8], segmentation method based on detection of local features [10-12], optical flow techniques [9]. It is well known that one good method of segmentation of objects that works efficiently with all the identified problems does not exist.

2. OBJECT EXTRACTION TECHNIQUE

In the classical segmentation algorithms, a major problem appears to be low quality video as well as problems resulting from the dynamically changing content of the images. Object segmentation algorithms do not calculate the position of the objects correctly, therefore, they are complemented by the object detection method basing on characteristic features search. For this purpose, a feature descriptor operating in an adaptively selected, predefined window around the position of the object (the window is a potential candidate to detect the object) is constructed. The best results are obtained using locally one of the methods: SIFT, SURF or HOG [12-14]. Generally, the idea of operation of these methods is similar and bases on finding stable local features within a defined search window. Features detected by the algorithms are chosen in such a way that they are not sensitive to changes in scale and orientation, as well as minor changes in illumination, noise, and shifting points of view. An important feature of these methods is resistance to partial covering of the objects. Therefore, these descriptors have become extensively used in the segmentation process improvement issue. The authors conducted a series of studies exploring the use of the HOG descriptor's properties [15-16].

Therefore, after analyzing various segmentation and tracking techniques, the authors proposed a solution that combines a segmentation method and a method of tracking of segmented regions using nonlinear classifiers and detec-

tor overrides. The proposed segmentation algorithm is shown in Figure 1.

The camera global motion estimation algorithm is used to improve scene objects segmentation and tracking algorithms and to detect a zoom in the analyzed sequence. This can also help to better adjust the size of detection windows used for object detection. The proposed algorithm is presented in Section 3.

The knowledge of the world coordinate system (obtained from camera calibration parameters) may also provide an opportunity to locate objects in the real-world coordinate system and specify their true positions in the scene. This may be helpful in object tracking applications, especially during the camera shot change, when the information about object positions obtained from different cameras must be combined together. The proposed algorithm is presented in Section 4.

The results achieved in the proposed solution for object segmentation in a case of sport sequences are presented in the Section 5.

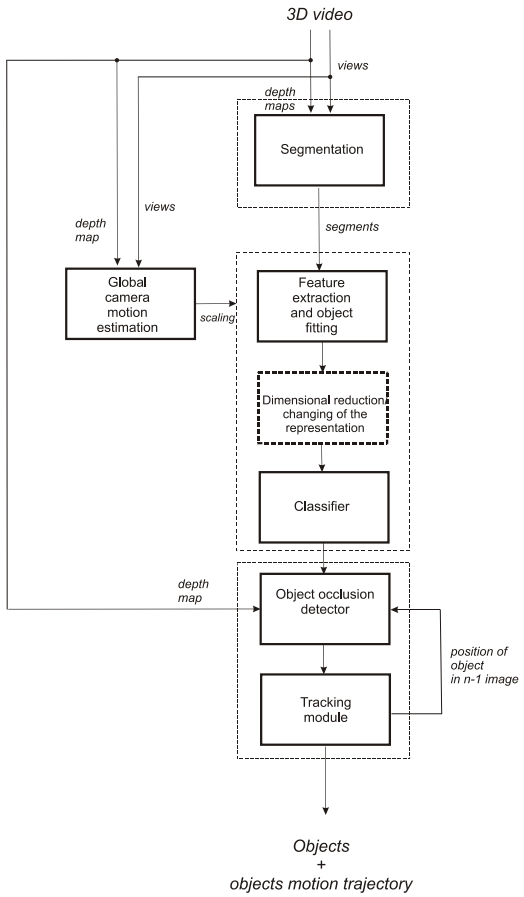

Fig.1 The proposed object extraction algorithm when we deal with multiple cameras, changing the camera position and focal length.

3. GLOBAL CAMERA MOTION ESTIMATION

The algorithm is based on the observation that only characteristic points of the image can be used to estimate the camera motion correctly. On this assumption, the following algorithm is proposed (Fig. 2):

- Harris corner detector – in order to select image points with good tracking ability (corners or clear texture objects, clear texture – sharp and non-noised texture).
- Discard feature points which do not belong to important areas.
- Calculation of optical flow for feature points selected in second step using hierarchical Lucas-Kanade inverse compositional algorithm and assuming translational motion model. In this approach 3 scales are used: ¼, ½ and full original image resolution to handle large inter-frame motions. As a result pairs of corresponding feature point coordinates for time instance t and $t+1$ are obtained:
  $[(x_1,y_1),(x_2,y_2),\ldots,(x_N,y_N)]_t \rightarrow$
  $[(x_1',y_1'),(x_2',y_2'),\ldots,(x_N',y_N')]_{t+1}$.
- Assuming affine warp global motion model with 6 parameters p_1, p_2, p_3, p_4, p_5, p_6 (describing translation, rotation and zoom):

$$\begin{cases} x' = (1+p_1)x + p_3 y + p_5 \\ y' = p_2 x + (1+p_4)y + p_6 \end{cases}$$

the affine warp parameters using the Least Squares Analysis can be calculated.

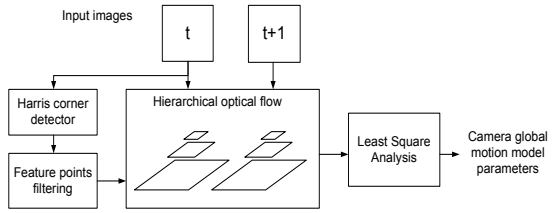

Fig. 2. Camera global motion estimation model.

4. CALCULATION OF CAMERA PARAMETERS FROM HOMOGRAPHY MATRIX

The translation and rotation parameters can be directly used to classify the type of the camera view and the area of 3D scene which is currently presented. This may be done with the predefined classification of camera view based on the rotation and translation the parameter sets characteristic for each class of view. What is noteworthy, this procedure provides much more accurate information than a simple analysis of scene area location based on the dominant color detection or color segmentation of the input image.

Another important application of the camera parameters set is the ability to track the parameter values in order to accurately determine camera zoom and panning.

Objects model fitting is used to determine a homography matrix H which transforms one plane (object model) into another plane (object observed in analyzed image). The

H matrix is an eight-parameter perspective transformation, transforming a point in the model coordinate system *(x', y', w')* into image coordinates *(x, y, w)*:

$$\begin{pmatrix} x \\ y \\ w \end{pmatrix} = \begin{pmatrix} h_{00} & h_{01} & h_{02} \\ h_{10} & h_{11} & h_{12} \\ h_{20} & h_{21} & h_{22} \end{pmatrix} \begin{pmatrix} x' \\ y' \\ w' \end{pmatrix}$$

where h_{22} is equal 1. The H matrix can be decomposed into internal camera parameters matrix, camera rotation and translation matrix and non-isotropic scaling matrix:

$$h = \begin{pmatrix} f & 0 & o_x \\ 0 & f & o_y \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} r_{00} & r_{01}\, r_{02} & t_x \\ r_{10} & r_{11}\, r_{12} & t_y \\ r_{20} & r_{21}\, r_{22} & t_z \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & B & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

If one compensates the camera principal point (o_x, o_y), assumed to be at (*image_width*/2, *image_height*/2), the matrix H turns into H' which can be decomposed into the following form [2]:

$$\begin{pmatrix} 1 & 0 & -o_x \\ 0 & 1 & -o_y \\ 0 & 0 & 1 \end{pmatrix} H = H'$$

$$H' = \begin{pmatrix} f & 0 & 0 \\ 0 & f & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} r_{00} & r_{01} & t_x \\ r_{10} & r_{11} & t_y \\ r_{20} & r_{21} & t_z \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 \\ 0 & B & 0 \\ 0 & 0 & 1 \end{pmatrix} = \begin{pmatrix} fr_{00} & Bfr_{01} & ft_x \\ fr_{10} & Bfr_{11} & ft_y \\ r_{20} & Br_{21} & t_z \end{pmatrix}$$

Since only non-isotropic scaling (or almost non-isotropic due to numerically sensitive computations of matrix H) is considered, the values of focal length f and non-isotropic scaling B from matrix H' can be easily obtained and, consequently, the rotation and translation parameters $\{r_{ij}\}$ and (t_x, t_y, t_z) can be calculated:

$$f^2 = \frac{h'_{00}h'_{01} + h'_{10}h'_{11}}{h'_{20}h'_{21}}, B^2 = \frac{{h'_{01}}^2 + {h'_{11}}^2 + f^2{h'_{21}}^2}{{h'_{00}}^2 + {h'_{10}}^2 + f^2{h'_{20}}^2}$$

To obtain values of rotation angles in 3D space from matrix H' the following rotation definitions are used:

$$R_x(\alpha) = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & \cos\alpha & -\sin\alpha & 0 \\ 0 & \sin\alpha & \cos\alpha & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

$$R_y(\beta) = \begin{pmatrix} \cos\beta & 0 & \sin\beta & 0 \\ 0 & 1 & 0 & 0 \\ -\sin\beta & 0 & \cos\beta & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

$$R_z(\gamma) = \begin{pmatrix} \cos\gamma & -\sin\gamma & 0 & 0 \\ \sin\gamma & \cos\gamma & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

Where $R_x(\alpha)$ rotates the *y*-axis towards the *z*-axis with angle α, $R_y(\beta)$ rotates the *z*-axis towards the *x*-axis with angle β and $R_z(\gamma)$ rotates the *x*-axis towards the *y*-axis with angle γ. The result of these three R matrixes differs depending on the order of the transformations, however, it may always be presented in the form of:

$$R = \begin{pmatrix} r_{00} & r_{01} & r_{02} & 0 \\ r_{10} & r_{11} & r_{12} & 0 \\ r_{20} & r_{21} & r_{22} & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

Matrix H' contains only $\{r_{00}, r_{01}, r_{10}, r_{11}, r_{20}, r_{21},\}$ values, therefore not all transformations may be easily computed from this matrix. For this implementation the order $R = R_z(\gamma)R_x(\alpha)R_y(\beta)$ was used:

$$R = R_z(\gamma)R_x(\alpha)R_y(\beta) = \begin{pmatrix} \cos\beta\cos\gamma - \sin\alpha\sin\beta\sin\gamma & -\cos\alpha\sin\gamma & \sin\beta\cos\gamma + \sin\alpha\cos\beta\sin\gamma & 0 \\ \cos\beta\sin\gamma + \sin\alpha\sin\beta\cos\gamma & \cos\alpha\cos\gamma & \sin\beta\sin\gamma - \sin\alpha\cos\beta\cos\gamma & 0 \\ -\cos\alpha\sin\beta & \sin\alpha & \cos\alpha\cos\beta & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

from which angles $\{\alpha, \beta, \gamma\}$ can be easily obtained.

Unfortunately, the camera zoom cannot be directly obtained from the matrix H'. However, as one may notice, the product of parameters f and B may be successfully used as its estimate:

$$zoom = fB$$

Using this equations we can determine the expected position of objects in the subsequent frames. This information is necessary to properly make the segmentation of objects in case of changing of the camera position and focal length.

The above equations have been introduced by the author and described in detail, since they were not previously available in literature in the form that would directly enable presentation of the camera rotation angles. It is one of the possible derivations which includes the value of rotation angle in the matrix formula.

5. EXPERIMENTAL RESULTS

In this section an evaluation of the proposed football player detection system is presented. In order to inspect the analyzed system in reference to different input video resolution, performance of the system should be evaluated using both SD (720×576) and HD (1280×720) test sequence resolution. Additionally, the real football test sequences consist of a video material with different lighting conditions, such as non-uniform playfield lighting, multiple player shadows etc., as well as different position of cameras used for the video acquisition. Taking the above conditions into consideration, 9 test sequences with football events and length of 25 to 50 frames were selected to form a test set for the system evaluation. For each test sequence the ground truth regions indicating football players were manually selected to create the desired output of the system. Presented system is evaluated with the precision and recall performance metrics, defined as follows:

$$precision = \frac{TP}{TP + FP}, \quad recall = \frac{TP}{TP + FN}$$

where *TP* is the set of true positives (correct detections), *FP* - the set of false positives (false detections) and *FN* - the set of false negatives (missed objects).

T determines a threshold which defines the degree of overlap required to recognize two regions as overlapping the same area of the analyzed image.

In order to evaluate the player detection system presented in this paper we use threshold *T* value equal 0.3. This value was chosen based on the interpretation of the *T* parameter: *T*=0.3 means that more than a half of the areas of a bounding box representing the detected object position and the ground truth box selected manually for this object overlap. This value is sufficient to evaluate how efficiently the objects are detected by the system. However, in case of evaluation of a system dedicated for the object segmentation and it's extraction from the background, higher *T* values should be applied.

Table 1 presents detailed evaluation results of our detection system with the reference to each test sequence included in the test set. The analysis of Table 1 show that selected test sequences provided diverse difficulty level for the player detection system, as the *precision* and *recall* metric values differ among the test set. However, the average *precision* value reached by the system is 0.98, and the average *recall* metric values do not exceed 0.89.

TABLE 1.
DETECTION SYSTEM EVALUATION RESULTS FOR THRESHOLD T=0.3 (P-*PRECISION*, R-*RECALL*).

Test video sequence	P	R
1. fast camera pan, uniform lighting	1.00	0.87
2. non-uniform lighting, shadows, occlusion	0.98	0.92
3. occlusion, various player's poses	1.00	0.89
4. non-uniform lighting, occlusion	0.98	0.94
5. non-uniform lighting, occlusion	1.00	0.89
6. interlaced, uniform lighting	1.00	1.00
7. motion blur, small figures, occlusion	0.97	0.82
8. motion blur, occlusion, uniform lighting	0.96	0.71
9. green uniforms, occlusion	0.97	0.97
Average	**0.98**	**0.89**

6. ACKNOWLEDGEMENT

This work was supported by the public funds as a research project.

CONCLUSION

In the paper, a novel approach to football broadcast video segmentation is proposed based on a combination of several techniques used to detect players: HOG descriptor detection and SVM classification which show potential robustness in case of great inconstancy of weather, lighting and quality of video sequences, global camera motion estimation, camera parameters estimation/3D coordinates of the space). Results show that proposed solution seems to achieve high objective and subjective notes in terms of precise location of the detected objects. Consequently, there are some works deserving further research in the proposed approach.

REFERENCES

[1] Izquierdo, E., "Disparity/Segmentation Analysis: Matching with an Adaptive Window and Depth-Driven Segmentation", IEEE Trans. Circuits Systems Video Techn., vol. 9, pp. 589-607, 1999.

[2] Markovic D. and Gelautz M., "Video object segmentation using stereo-derived depth maps", 27th Wokshop of the AAPR/ÖAGM, pages 197-204, Laxenburg, 2003.

[3] Lucena M.J., Fuertes J.M., Gomez J.I., de la Blanca N.P., Garrido A., "Tracking from optical flow", Image and Signal Processing and Analysis, 2003, ISPA 2003, Proceedings of the 3rd International Symposium on Volume 2, 18-20 Sept. 2003 Page(s):651 - 655 Vol.2.

[4] Chung-Ming K., Mao-Hsiung H., Chaur-Heh H. ; "Baseball Playfield Segmentation Using Adaptive Gaussian Mixture Models"; Innovative Computing Information and Control, 2008. ICICIC '08. 3rd International Conference on, 18-20 June 2008 Page(s):360 – 360.

[5] Chen-Yu C., Jia-Ching W., Jhing-Fa W., Yu-Hen H., "Event-Based Segmentation of Sports Video Using Motion Entropy", Ninth IEEE International Symposium on Multimedia, 10-12 Dec. 2007 Page(s):107 – 111.

[6] Nunez, J.R.; Facon, J.; de Souza Brito, A.; "Soccer video segmentation: Referee and player detection"; Systems, Signals and Image Processing, 2008. IWSSIP 2008. 15th International Conference on, 25-28 June 2008 Page(s):279 – 282.

[7] Grau, O.; Thomas, G.A.; Hilton, A.; Kilner, J.; Starck, J.; "A Robust Free-Viewpoint Video System for Sport Scenes", 3DTV Conference, 7-9 May 2007 Page(s):1 – 4.

[8] Jungong H., Farin, D., de With, P., "Broadcast Court-Net Sports Video Analysis Using Fast 3-D Camera Modeling"; Circuits and Systems for Video Technology, IEEE Transactions on; Volume 18, Issue 11, Nov. 2008 Page(s):1628 – 1638.

[9] Lifang W.,; Xianglong M., Xun L., Shiju C., "A New Method of Object Segmentation in the Basketball Videos", Pattern Recognition, 2006. ICPR 2006. 18th International Conference on Volume 1, 0-0 0 Page(s):319 – 322.

[10] Salembier, P. and F. Marques, Region-based Representations of Image and Video: Segmentation Tools for Multimedia Services, IEEE Transactions on Circuits and Systems for Video Technology, vol. 9, no. 8, pp. 1147- 1169, 1999.

[11] Xinguo Y., Xiaoying T., Ee Luang A., "Trajectory-Based Ball Detection and Tracking in Broadcast Soccer Video with the Aid of Camera Motion Recovery", IEEE International Conference on Multimedia and Expo, 2-5 July 2007 Page(s):1543 – 1546.

[12] Grabner M., Grabner H., Bischof H., "Fast Aproximated SIFT", Asian Conference on Computer Vision, Washington, 1999.

[13] Lowe D., "Object Recognition from Local Scale-Invariant Features", Conference on Computer Vision, 2004,

[14] Muja M., Lowe D., "Fast Approximate Nearest Neighbors with Automatic Algorithm Classification" VISAPP Internation Conference on Computer Vision Theory and Application, Lizbona 2009.

[15] Mackowiak S., Konieczny J., Kurc M., Mackowiak P., "Football Player Detection in Video Broadcast", Lecture Notes on Sience 2010 Computer Vision and Graphics: Proc. ICCVG 2010 Volume Editor(s): L. Bolc, R. Tadeusiewicz, L.J. Chmielewski.

[16] Mackowiak S., Konieczny J., Kurc M., Mackowiak P., „A complex system for football player detection in broadcasted video", ICSES 2010.