TOTAL OCCLUSION CORRECTION USING INVARIANT WAVELET FEATURES

Mohammed Ghazal and Aishy Amer

Electrical and Computer Engineering, Concordia University, Montréal, Québec, Canada
Email:{moha_mo,amer}@ece.concordia.ca

ABSTRACT

This paper proposes a method which utilizes invariant wavelet features for correcting total occlusion in video surveillance applications. The proposed method extracts invariant wavelet features from the pre-occlusion spatial image of disappearing objects. When new objects are detected during occlusion, their extracted invariant wavelet features are compared to those of lost objects to check for reappearance. When reappearance occurs, the proposed method rebuilds the correct correspondence map between pre-occlusion and post occlusion objects to continue to track the ones that were lost during total occlusion. Our results show that the proposed method is more robust than referenced methods especially when objects change or reverse their motion direction during occlusion.

Index Terms— Tracking, wavelet transforms, video signal processing.

1. INTRODUCTION

Video object tracking can be defined as the process of creating unique correspondences between objects in a video sequence. With such correspondences, high-level semantics such as events and behaviors of objects can be extracted. The tracking process is challenged by incidents of occlusion. Occlusion takes place when one or more objects partially or totally mask regions of others which is common in real video sequences.

Occlusion handling usually depends on the nature of the tracking method used which can be classified as template-based, layer-based, and feature-based. In template-based methods such as [1], the absence of the tracked template features from the frame indicates occlusion. Total occlusion is corrected when the template features reappear. The problem in [1] is that it must learn in advance the features of all tracked templates which is not suitable in surveillance. Layer-based methods such as [2] define a pixel-wise layer visibility measure and use it to correct occlusion by separating the scene into layers. The drawback in layer-based methods is the high computational cost associated with computing and maintaining the layer visibility information. Particle filters [3] have recently become a popular method for visual tracking which copes with partial and short-lived occlusions. However, in many applications, the prior information available for the environment is limited (e.g., tracker cannot be initialized with features of objects of interest). Moreover, the complexity of the tracking process increases with multiple objects and a single camera. Feature-based methods such as [4, 5] extract sets of features for the objects and build inter-frame correspondences between them based on feature similarity. The problem with such methods is the unpredictable behavior of objects during total occlusion which may alter their features.

Occlusion can sometimes lead to total disappearance of occluded objects from the scene. When objects reappear after the end of occlusion, the problem becomes how to recover their pre-occlusion correspondences or identifications in order to continue to track them successfully. The complexity of the problem increases when multiple objects are involved which can exhibit during the occlusion changes in position, motion, size and orientation. With only the spatial image of the object from before the occlusion available, there is need for spatial features which tolerate a certain degree of transformation.

In this paper, we enhance the method in [6] to extract robust invariant wavelet features and propose a method that uses these features to solve the total occlusion problem. The key contributions in this paper are; 1) successful recovery from total occlusion based only on the spatial texture of objects, and 2) robustness to significant motion changes. The remainder of the paper is as follows. Sections 2 presents the proposed approach. Simulation results are discussed in section 3 and section 4 concludes the paper

2. PROPOSED METHOD

In this paper, the tracking method in [4] is used to perform the object detection and tracking tasks. The terms *image object* and *video object* are used extensively throughout this paper. An *image object* identifies a closed contour in the current frame. There is no temporal information associated with image objects as they live only in the current frame F_k. On the other hand, a *video object* refers to a temporally consistent object with temporal information such as motion, trajectory, or visibility. A *video object* is updated at the end of the tracking process with the information of the matched *image object*.

Let I_i denotes the i^{th} image object in the current frame

This work was supported, in part, by the Fonds de la recherche sur la nature et les technologies du Québec (NATEQ).

F_k and V_j as the j^{th} video object in the previous frame F_{k-1}. The tracking algorithm in [4] builds correspondences M_{ij} between I_i in F_k and V_j in F_{k-1} based on similarities in position, shape, size, and motion. Feature similarities are affected by occlusion. Fig 1 shows the different stages of a total occlusion incident when one image object is nearly invisible from the scene because it is masked by another image object. The tracker is challenged by the object reappearance in Fig 1(c) because: 1) the new position of object 2 is closer to the last known position of object 1, 2) The size features such as width, height, and area are very similar between the two objects; a situation which most likely continues to be true if more objects are involved, 3) The motion features are unreliable because the objects can reverse or change direction or speed during occlusion, and 4) The shape features are unreliable because the objects can shrink, expand or deform during occlusion. In the proposed method, we use invariant wavelet features because they rely on texture which changes less from before and after occlusion than other spatial features.

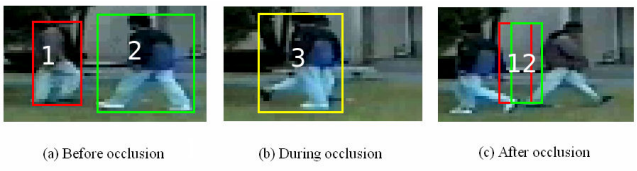

Fig. 1. Effect of occlusion in confusing a video object tracker

Once occlusion is detected, the invariant wavelet features are extracted based on [6] with proposed preprocessing steps and improved feature selection. The features are kept as part of the video object features. If a new image object is detected within a predefined search radius of the visible occluding object, the proposed method compares the features of the new image object with the stored features for all occluded video objects before declaring the image object as a new video object. If the features of the new image object are similar within a predefined threshold to the features of the occluded video object with minimum feature difference among all occluded video objects, the new image object is considered the reappearance of that occluded video object.

2.1. Occlusion Correction

Fig. 2 shows a block diagram of the proposed occlusion correction method. The first step of the proposed method is to detect all occluded objects between F_{k-1} and F_k using the occlusion detection method in [4] and store them in $\mathbf{V_{occ}}$, the set of occluded video objects. Next, for every $V_j \in \mathbf{V_{occ}}$, we create a new bounding box BB_{V_j} using

$$\text{BB}_{V_j}^{width} = max(\text{MBB}_{V_j}^{width}, \text{MBB}_{V_j}^{height}). \quad (1)$$

$$\text{BB}_{V_j}^{height} = max(\text{MBB}_{V_j}^{width}, \text{MBB}_{V_j}^{height}). \quad (2)$$

$$\text{BB}_{V_j}^{center} = \text{MBB}_{V_j}^{center} \quad (3)$$

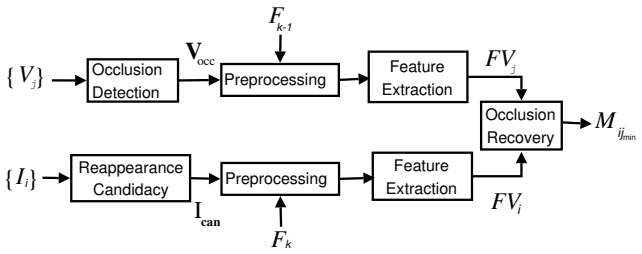

Fig. 2. Block diagram of proposed method.

where $\text{MBB}_{V_j}^{width}$ is the width of the minimum bounding box of V_j, $\text{MBB}_{V_j}^{height}$ is its height, and $\text{MBB}_{V_j}^{center}$ is its center. We do this to get a square image of the object required by the subsequent transformations.

The next step is to nullify the background texture effect as it will be different from before and after the occlusion. To do that, let $F_k(x, y)$ denote the pixel at spatial position (x, y) and frame k of size $M \times N$. We set to zero all pixels in BB_{V_j} and not in V_j using contour filling (see Fig. 3), i.e.,

$$F_{k-1}(x, y) = 0, \quad \forall (x, y) \in \text{BB}_{V_j} - V_j, \quad (4)$$

Note that we apply contour filling on the binary image of the

Fig. 3. Expansion of the MBB of the object and nullifying the effect of the background texture.

object from background subtraction, then apply the result to the colored image. After the proposed previous preprocessing steps, we extract the invariant wavelet features of the sub-image $F_{k-1}(x, y)$ $(x, y) \in \text{BB}_{V_j}$ using [6] to obtain the feature vector $\mathbf{FV}_j$ for V_j and store it as part of its information. The reason we use invariant wavelet featuers is to tolerate a degree of deformation in the object shape from one frame to another. The process of extracting the invariant wavelet features begins by transforming the object's square gray-level sub-image $F_{k-1}(x, y)$, $(x, y) \in \text{BB}_{V_j}$ using the log-polar transform to obtain an $S \times R$ log-polar image $lp(u, o)$. Then, we apply the Discrete Wavelet Packet Transform (DWPT) to $lp(u, o)$ and its one-row circular shift down version to create an oct-tree. Formally, let i denote the sub-band index (e.g., $i \in \{LL, LH, HL, LL\}$), j denote the decomposition level, ϕ denote the scaling function, and φ denote the wavelet function, the DWPT is done by

$$W_\phi^i(j, m, n) = \frac{1}{\sqrt{MN}} \sum_{m=1}^{M-1} \sum_{n=1}^{N-1} lp(u, o) \phi_{j,m,n}^i(u, o), \quad (5)$$

$$W_\varphi^i(j, m, n) = \frac{1}{\sqrt{MN}} \sum_{m=1}^{M-1} \sum_{n=1}^{N-1} lp(u, o) \varphi_{j,m,n}^i(u, o), \quad (6)$$

where $W_\phi^i(j,m,n)$ denotes the approximation wavelet coefficients at level j of sub-band i, and $W_\varphi^i(j,m,n)$ denotes the details wavelet coefficients at level j of sub-band i. The octtree is adaptively pruned based on a information cost (IC) function in order to decrease the number of computations. The tree is pruned as in [6]. We use the IC function

$$IC(W_\phi^i(j,m,n)) = \sum_{m,n} ln(W_\phi^i(j,m,n)^2). \quad (7)$$

Finally, we compute energy signatures for the sub-band in the pruned tree and the L most dominant (largest) signatures are used as a feature vector. We use the energy signature function

$$ES(W_\phi^i(j,m,n)) = \frac{1}{MN}\sum_{m,n}|W_\phi^i(j,m,n)|. \quad (8)$$

We propose to improve the feature vector selection process by using only the energy signature of the approximation coefficients at the first level $j=1$ (e.g., $ES(W_\phi^{LL}(1,m,n))$) as part of the feature vector. The approximation coefficients at subsequent levels which come from approximation coefficients in previous levels are ignored because their energy signatures are similar and close to the signal mean.

As long as we have objects which are lost during occlusion, we extract the invariant wavelet features of all image objects I_i in $\mathbf{I_{can}}$, the set of image objects which are candidates to be the reappearance of lost video objects. $\mathbf{I_{can}}$ is populated with all image objects in F_k which do not have a correspondence with any video object in F_{k-1} and are within distance d (e.g., $d = \frac{1}{2}min(M,N)$ is adapted to the frame size and is large enough to account for fast moving objects) from the occluding object. Let $\mathbf{FV}_i$ denote the feature vectors of the candidate image objects. We first find the video object with the minimum feature difference using

$$j_{min} = \underset{j}{\text{argmin}}\ |\mathbf{FV}_i - \mathbf{FV}_j|,\ V_j \in \mathbf{V_{occ}}, \quad (9)$$

and then check if the difference is beyond a certain threshold Th for similarity (i.e., $|\mathbf{FV}_i - \mathbf{FV}_{j_{min}}| < Th$) ($Th$ is experimentally chosen, e.g., $Th = 1000$). If so, we declare the correspondence $M_{ij_{min}}$ to indicate the recovery.

2.2. Feature invariance of the proposed method

We can model the transformation of the pixels inside the object between frames using an affine model. Since objects are segmented and are matched based on features, the translation of objects is not important. We can then model the transformation of objects during occlusion as a combination of scaling and rotation. Because of circular sampling in the log-polar transform, the rotation effect is nullified within a row-shift. However, the scale invariance of the method is not as strong possibly due to the different visual content in scaled objects.

The log-polar transform expects a square image as input. The image of extracted objects from a video sequence can be non-square and their size not a power of two. We choose to expand the bounding box and remove the background instead of resizing the object's subimage because the later requires low-pass pre-filtering. The low-pass filtering smears details which are vital to capture the texture of the image object. Moreover, since this method is supposed to rely on variations or texture to identify objects, we proposed to not use energy signatures calculated from approximation coefficients. We refer here to the approximation sub-bands resulting from decomposing approximation and not details sub-band. The energy signature of such sub-bands are always part of the dominant signatures (large valued coefficients) without contributing to the inter-class differences. Moreover, when using the ES in (8), the values of the energy signatures of these sub-bands are very similar and close to the signal average.

3. RESULTS

To experimentally evaluate the proposed method, we use the four video sequences in Fig. 4. The video sequence in Fig. 4(a) shows five objects with different texture patterns and shapes at different stages of total occlusion (before occlusion $k=14$, during occlusion $k=16$, and at the end of occlusion $k=18$). The five objects occlude into one and exhibit during occlusion, rotation, scaling, and significant change of motion direction. Objects 1 and 4 reverse their motion directions and return to their original location. Four of the objects are lost during occlusion but are successfully recovered by the proposed method whereas the referenced method [5] recovers only object 3, declares objects 0 as new object 5, and confuses objects 1 and 4. This is because the proposed method relies on invariant texture features which allow for a degree of transformation during occlusion and do not rely on motion.

Fig. 4(b) shows a manually changed $survey$ sequence. The original objects in frame $k=91$ of the survey sequence only partially occlude. We change the multiple partial occlusion into a multiple total occlusion. The proposed method is able to recover all lost objects during occlusion in $k=113$, whereas the referenced method recovers only the object with the least amount of transformation.

Fig. 4(c) shows a natural total occlusion in the $comm2$ sequence. While the proposed method continues to assign the correct identifications after the end of occlusion, the referenced method [5] confuses the two object. Finally, Fig. 4(d) shows a different total occlusion scenario in the commonly referenced $ekrlb$ sequence as an object is occluded with a large obstacle (the column) and is lost from the scene in frame $k=49$. When, the object reappears from behind the obstacle in frame $k=60$, the proposed method is able to continue to track it while the referenced method [5] declares the reappearing object as a new object (object 1) and loses the original object (object 0). Also, the occlusion was prolonged by one minute and the same results were obtained. This is attributed to the proposed method refraining from declaring new objects with the presence of lost objects from the scene during occlusion until their features are checked for reappearance.

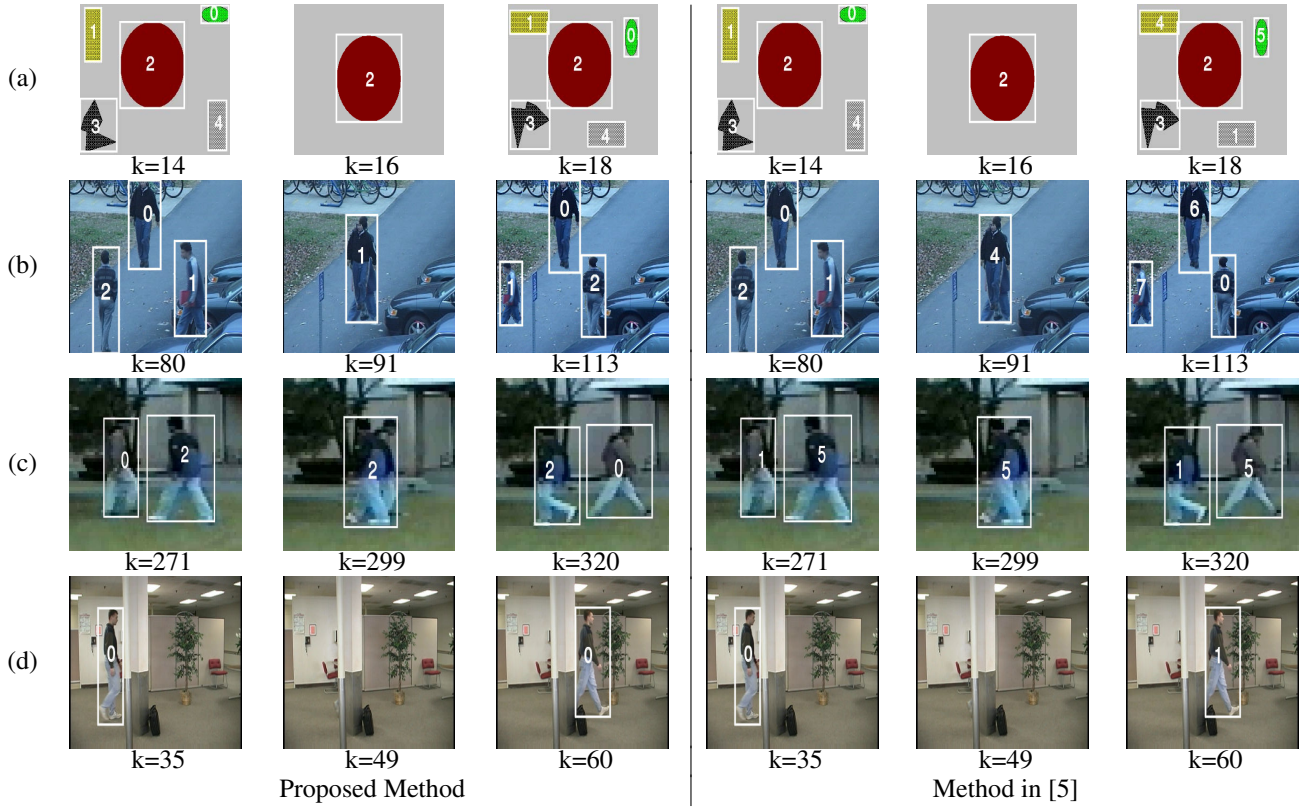

Fig. 4. The tracking results for proposed and referenced method[5] after recovery from total occlusion at different scenarios. Note that different trackers deal with small objects differently, hence the different labels between the proposed method and [5].

4. CONCLUSION

This paper proposed a method for correcting total occlusion in video surveillance applications using invariant wavelet features. The proposed method extracts texture-based invariant wavelet features from the gray-level image of disappearing objects during occlusion and uses them to check for reappearance. The extracted invariant wavelet features are based on a method which uses the energy signatures of the best basis representation of the DWPT of the log-polar transformed object images with proposed improved preprocessing and feature selection. The invariance to rotation in the extracted features is more robust than invariance to scaling. The correct correspondence is re-established after occlusion to continue to track lost objects. Our results show that the proposed method is more robust than referenced methods especially when objects change or reverse their motion direction during occlusion.

5. REFERENCES

[1] D. J. Bullock and J. S. Zelek, "Real-time tracking for visual interface applications in cluttered and occluding situations," *Image and Vision Computing*, vol. 22, pp. 1083–1097, Oct. 2004.

[2] Y. Zhou and H. Tao, "A background layer model for object tracking through occlusion," in *Ninth IEEE Int. Conf. on Comp. Vision*, Oct. 2003, vol. 2, pp. 1079–1085.

[3] J. Rittscher, N. Krahnstoever, and L. Galup, "Multi-target tracking using hybrid particle filtering,," *Seventh IEEE Workshops on Application of Computer Vision WACV/MOTION05*, vol. 13, pp. 447–454, 2005.

[4] Aishy Amer, "Voting-based simultaneous tracking of multiple video objects," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 15, pp. 1448–1462, Nov. 2005.

[5] T. Yang, Q. Pan, J. Li, and S. Z. Li, "Real-time multiple objects tracking with occlusion handling in dynamic scenes," in *Proc. IEEE Conf. Computer Vision Pattern Recognition*, June 2005, vol. 1, pp. 20–25.

[6] C. M. Pun and M. C. Lee, "Log-polar wavelet energy signatures for rotation and scale invariant texture classification," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 25, pp. 590–603, May 2003.