Les Moments de Zernike spatio-temporels pour la description et la classification des actions dans les vidéos

Imen LASSOUED*, Ezzeddine ZAGROUBA*,
Youssef CHAHIR**

*Equipe de recherche SIIVA
laboratoire RIADI
Université de Tunis El manar
Institut Supérieur d'informatique (ISI)
**Laboratoire GREYC, Groupe de Recherche Image
Université de Caen

Résumé. La reconnaissance et la classification des actions dans les images et les vidéos est l'un des sujets de recherche d'actualité dans le domaine de reconnaissance des formes et de vision par ordinateur. Cet article propose une nouvelle méthode de classification des actions dans les vidéos basée sur les moments de Zernike spatio-temporels. L'application de ces moments permet de capturer à la fois l'information structurelle et temporelle d'une séquence vidéo. La première étape de cette méthode consiste à segmenter la vidéo en un volume spatio-temporel d'images. Ensuite, l'ensemble de silhouettes des objets en mouvement sera extrait à partir de ces images, cet ensemble définira une forme spatio-temporelle. L'étape suivante consiste à appliquer les moments de Zernike spatio-temporels sur la forme obtenue. Cette démarche permet de définir un descripteur pour chaque vidéo de la base de données. Ces descripteurs serviront ensuite à classifier les actions de ces vidéos en utilisant la méthode des moindres carrés de l'algorithme SVM. L'expérimentation de notre méthode sur les deux bases de vidéos d'actions Weizmann et KTH a donné des résultats efficaces en terme de caractérisation et de classification d'actions.

1 Introduction

Les volumes des contenus audio-visuels mis à disposition des utilisateurs ne cessent de croître. l'indexation et la recherche des vidéos par l'analyse automatique de leurs contenu constitue l'un des enjeux majeurs des systèmes d'information. Dans ce contexte, il est important d'extraire automatiquement les informations de haut niveau qui peuvent décrire le contenue sémantique de la vidéo proposée. En effet, plusieurs travaux introduisent la notion d'actions en vidéo pour plusieurs applications telles que : l'analyse sportive, la surveillance visuelle et l'interaction homme-machine. La classification d'actions consiste à classer les vidéos en se basant sur l'action de l'objet détecté. Dans cet article, nous présentons une nouvelle approche de classification des actions dans les vidéos basée sur les moments de Zernike spatio-temprels. Ces moments sont calculés en utilisant les volumes de silhouettes. Cet article est organisé

comme suit : la première section présente un aperçu des méthodes actuelles de reconnaissance et classification d'actions. Dans la deuxième section, nous détaillons l'approche proposée qui est basée sur les moments de Zernike spatio-temporels et le classificateur LS-SVM. Les résultats expérimentaux et les évaluations sont présentées dans la section 3. On finira par une conclusion et des perspectives dans la section 4.

2 Etat de l'art

Les approches existantes de reconnaissance et de classification des actions dans les vidéos peuvent être classées en trois principales catégories basées sur les descripteurs utilisés.

2.1 Méthodes basées sur le flot optique, le gradient et l'intensité

Nombreux travaux en reconnaissance et classification d'actions sont fondés sur les caractéristiques globales de la vidéo telle que le flux optique, l'histogramme de gradient et l'intensité. Zelnik-Manor et al(2001) ont utilisé des histogrammes de gradient spatio-temporels avec des échelles temporels multiples pour classer les actions dans les vidéos. Wu(2005) à développé un algorithme pour extraire une fenêtre centrée sur la forme suivie. Il propose aussi de décomposer le gradient de l'image en quatre canaux pour classifier les actions des personnes. Efros et al(2003) ont proposé un descripteur basé sur des mesures floues de flux optique et il l'a appliqué pour reconnaître les actions dans le ballet, le tennis et le jeu de football. Dollar et al(2005)ont proposé de caractériser les actions en utilisant les points d'intérêts spatio-temporels extraits des vidéos. Dans cette approche, l'action est décrite selon le type et l'emplacement des points d'intérêt détectés dans la séquence. Pour cette classe de méthode, les résultats de classification dépendent beaucoup des conditions d'enregistrement de la vidéo.

2.2 Méthodes basées sur le suivi de mouvement

De nombreuses méthodes de classification d'actions sont fondées sur le suivi de l'objet, soit dans un espace 2D ou 3D (voir Gavrilla (1999)) ou encore (Cedras et al (1995)) . Rao et Shah (2001) ont proposé une approche basée sur la trajectoire des mains pour différencier les actions. Cedras et al (1995) ont utilisé un arrangement spatial des points d'intérêts suivis pour distinguer entre les actions " walking " et " biking ". Song et al (2003) ont proposé de représenter les actions en utilisant les courbes obtenues à partir des résultats de suivi de cinq parties du corps. Yacoob et black(1999)ont utilisé l'information 3D pour établir des descripteurs de mouvement basés sur les positions, les angles et les vitesses des parties du corps. Ali et Aggarwal(2001) ont utilisé les angles d'inclinaison du torse, les parties inférieures et supérieures des jambes comme des caractéristiques pour reconnaître l'action et l'activité. Dans ces approches, on remarque que la tache de suivi pourrait être complexe en raison de la grande variabilité dans la forme et l'articulation du corps humain.

2.3 Méthodes basées sur les silhouettes des objets

Actuellement, plusieurs travaux de recherche dans la classification d'actions sont orientés vers des méthodes à base de silhouettes (voir Bobick et Davis(2001)) ou encore (Weinland

et al (2005)). En effet, les actions de l'homme peuvent être caractérisées par le mouvement dynamique d'une séquence de silhouettes humaines. Gorelick et al(2007) ont utilisé le volume spatio-temporel des silhouettes pour la caractérisation de l'action. En effet, il a calculé les propriétés de la solution de l'équation de Poisson pour extraire les caractéristiques spatio-temporelles des silhouettes. Guo et al (2009) ont considéré l'action comme une séquence temporelle des déformations locales de la silhouette de l'objet .Il ont représenté chaque action par la matrice de covariance empirique d'un ensemble de treize dimensions normalisées de vecteurs géométriques capturant la forme de silhouette. Sminchisescu et al(2005)ont caractérisé l'action de l'homme sur la base discriminatoire de champ aléatoire conditionnel (CRF) et de l'entropie maximale des modèles de Markov (MEMM). Il a utilisé des descripteurs d'images combinant le contexte de la forme et les caractéristiques de points d'intérêts extraites sur la silhouette. L'avantage de cette classe de méthodes est l'extraction aisée de silhouettes humaines pour les techniques de vision actuelles en particulier dans l'imagerie où les caméras sont fixes. Nous présentons dans ce papier une nouvelle méthode de classification d'actions qui se base sur les silhouettes des objets en mouvement.

3 Approche proposée

L'objectif principal de l'approche proposée est de classifier les actions dans les vidéos. Notre contribution consiste à étendre le modèle des moments de Zernike 2D dans le domaine spatio-temporel et d'utiliser leurs fonctionnalités pour classer les actions des personnes. Dans notre approche, l'action est représentée par le volume spatio-temporel des silhouettes de l'objet en mouvement. Ce volume de silhouette est considéré comme une forme 3D contenant à la fois l'information spatiale sur la pose de la personne et l'information liée à sa dynamique de mouvement. Notre approche consiste à effectuer une segmentation temporelle de la vidéo afin d'obtenir un volume d'image. Ensuite, l'arrière plan de cet ensemble d'images est extrais pour obtenir le volume des silhouettes. Puis, on calcule les moments de Zernike spatio-temporels pour chaque série de silhouettes obtenue. Enfin, la méthode de classification LS-SVM est appliquée pour réaliser une classification des vidéos de la base de données en entrée. FIG.1 présente les différentes étapes de l'approche proposée.

3.1 Extraction de volume spatio-temporel des silhouettes

La première étape de notre approche est l'extraction de volume spatio-temporel des silhouettes de la personne effectuant une action dans la vidéo. Nous détectons les personnes se déplaçant par une soustraction du fond. Nous utilisons le dictionnaire de soustraction de fond décrit dans fihl et al(2006) et dans Kim et al(2005)). Ce dictionnaire gère les ombres et le camouflage du premier plan en séparant l'intensité et la chromaticité dans le modèle de fond. Le modèle conçu est multi modale ce qui permet de modéliser le mouvement de fond comme l'exemple des branches d'arbres. fihl et al(2006) décrivent deux mécanismes différents qui gèrent la mise à jour rapide et progressive des changements pour maintenir le modèle de base cohérent à tout moment. Cette méthode robuste de soustraction de fond nous permet d'utiliser des séquences vidéo assez diverses avec plusieurs scénarios pour mettre en évidence notre méthode de classification d'actions.

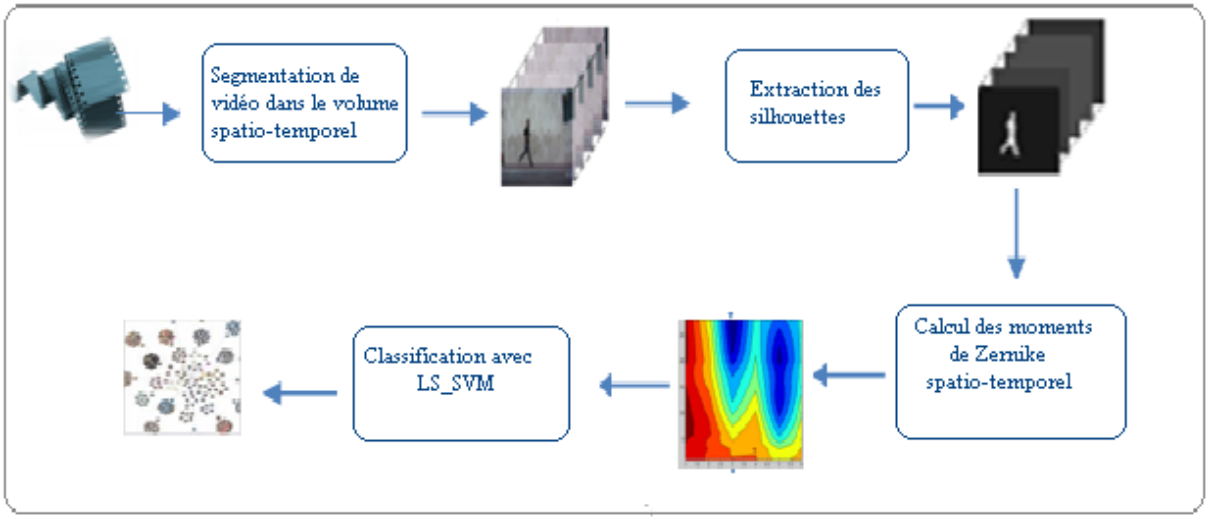

FIG. 1 – *Architecture générale de l'approche proposée*

3.2 Calcul des moments de Zernike spatio-temporels

La principale contribution de notre approche consiste à utiliser une extension des moments de Zernike dans le domaine spatio-temporel pour la classification des actions dans les vidéos. En effet, les moments de Zernike sont considérés parmi les moments invariants les plus efficaces en terme de performance globale pour la représentation et reconstruction des formes. En plus, ces moments sont invariants par rotation et peuvent être facilement reconstruits à un ordre arbitraire. Les moments de zernike sont calculés selon un ordre bien définit. En effet, les moments d'ordres supérieurs offrent une caractérisation fine pour une forme donnée, mais peuvent devenir plus sensibles au bruit. Shutter et Nixon (2001) ont montré que les moments de Zernike présentent un bon taux de reconnaissance et une description compacte lorsqu'ils sont appliqués pour analyser des séquences de marche. Dans l'approche proposée, les moments de Zernike spatio-temporels sont choisis pour caractériser les actions dans des séquences vidéos. Par conséquent, les moments sont calculés pour chaque série de silhouettes et expérimentés avec différents ordres afin de déterminer un ordre optimal capable de résoudre le problème. les moments de Zernike peuvent être calculés par deux méthodes. La première méthode consiste à confondre l'axe temporel à l'axe spatial Z, ainsi on obtient des blocs d'images représentés en trois dimensions XYT (x, y et le temps). En conséquence, l'information temps est confondu avec l'information espace.

La deuxième méthode traite le temps séparément de l'espace parce qu'ils sont fondamentalement différents. Pour notre travail, nous avons choisi d'appliquer la deuxième méthode car elle permet de reformuler les moments en intégrant le paramètre temps. Cette méthode permet de séparer les descriptions spatiales et temporelles.

Les formes en mouvement se divisent en formes rigides et non-rigides. Les moments des formes rigides en mouvement décrivent simplement l'information spatiale puisque les images

sont fortement corrélées. Dans ce cas, il suffit d'affiner la description de la forme rigide et on pourra ainsi diminuer la longueur de la séquence. Les moments de Zernike de cette séquence peuvent être considérés comme la moyenne des moments de Zernike d'une seule image.

Dans le cas ou les formes sont non-rigide comme l'exemple d'une personne qui se déplace, la corrélation spatiale entre les descriptions des images consécutives est réduite. Le moment de Zernike de la séquence en question est composé de la somme des moments de Zernike des images pondérée par la vitesse. On rappelle que notre application s'intéresse aux formes non rigides. Les descripteurs finaux des formes non rigides en mouvement sont corrélés temporellement en raison de l'utilisation de la séquence d'images. Nous avons utilisé les expressions des moments de Zernike 2D issues de Teague(1979) et nous avons effectué une extension de ces expressions dans le domaine spatio-temporel.Ainsi les moments de Zernike spatio-temporels noté A_{mn}, sont exprimés par :

$$A_{mn} = \frac{m+1}{\pi} \sum_{i=2}^{images} \sum_{x} \sum_{y} U(i,\mu,\gamma) S(m,n) P_{ixy} \tag{1}$$

Avec n c'est l'ordre de moment et m est un entier tel que $n - |m| \geq 0$ et P_{ixy} est la valeur du pixel (x,y) dans l'image i.

L'expression des polynômes orthogonaux S_{mn} est la suivante :

$$S(m,n) = [V_{mn}(r,\theta)]^* \tag{2}$$

où * désigne le conjugué du complexe.Les $V_{mn}(x,y)$ sont les polynômes de Zernike et sont exprimés en coordonnées polaires par :

$$V_{mn}(r,\theta) = R_{mn}(r) exp(jn\theta) \tag{3}$$

où $(r;\theta)$ sont définies sur le cercle d'unité et $R_{mn}(r)$ est le polynôme orthogonal radiale, définie par :

$$R_{mn} = \sum_{s=0}^{\frac{m-|n|}{2}} (-1)^s F(m,n,s,r) \tag{4}$$

Avec

$$F(m,n,s,r) = \frac{(m-s)!}{s!(\frac{m+|n|}{2}-s)!(\frac{m-|n|}{2}-s)!} \tag{5}$$

3.3 Classification basée sur LS-SVM

Plusieurs travaux ont été effectués pour classer les actions dans les vidéos en utilisant différentes approches telles que les réseaux de neurones (voir Baccouche et al (2010)), GMM (voir liangliang et al(2001)), etc. Le classificateur le plus utilisé dans les travaux de cette thématique est la machine SVM qui permet d'obtenir des résultats efficaces dans de nombreux problèmes de reconnaissance des formes et d'actions. Pour cette raison, nous avons choisi d'utiliser la variante LS-SVM de SVM qui simplifie la formulation de SVM sans aucune perte de ses avantages.

La variante LS-SVM a été proposée par Suykens et Vandewalle (2000) comme un algorithme d'apprentissage permettant de résoudre un problème convexe . Il a été démontré par une étude minutieuse empiriques que les performances de généralisation de LS-SVM est comparable à celle de SVM (voir van gestel et al (2000)). En outre, l'algorithme d'apprentissage de LS-SVM est très simplifié puisque un problème linéaire est résolu au lieu d'un problème de programmation quadratique (QP) dans le cas de SVM. Étant donné un ensemble de formation $(x, y), i = 1, 2, , l$ avec des données d'entrée $x_i \in R^n$ et les données de sortie $y_i \in R^n$. Le classificateur prend la forme suivante :

$$y(x) = sign[w^T \varphi(x) + b] \tag{6}$$

Avec $\varphi(x) : R^n \rightarrow R^n$ est une transformation d'espace. Le problème d'optimisation devient

$$\min_{w,b,e} J_p(w, e) = \frac{1}{2} w^T w + \gamma \frac{1}{2} \sum_{i=1}^{N} e_i^2 \tag{7}$$

$$y_i[w^T \varphi(x_i + b] = 1 - \xi_i, i = 1, 2, .., l \tag{8}$$

$\xi_i > 0$ désigne une constante réelle utilisée pour contrôler l'erreurs de classification.

La référence Suykens et Vandewalle (2002) donne des explications plus détaillées pour le calcul de la variante LS-SVM.

4 Résultats expérimentaux

4.1 Base de données d'actions Weizmann

Nous allons utiliser la base de données accessible au public proposée par Blenck et al (2005) pour évaluer notre approche . Cette base est généralement utilisée pour tester des modèles de classification d'actions. Elle contient 81 vidéos avec une basse résolution (180 * 144). Ces vidéos contiennent 9 personnes qui effectuent 9 actions différentes (running, bending, waving with one hand, jumping in place, jumping jack, jumping, walking, skip, and waving with two hands). Chaque séquence vidéo contient un objet effectuant une action unique. Des exemples pour chacune de ces actions sont présentés dans FIG.4. Le calcul des moments de Zernike nécessite un prétraitement des vidéos qui consiste à segmenter la vidéo en un volume spatio-temporel et extraire les silhouettes associées aux images. On rappelle que les moments se calculent uniquement pour des images carrées. Ceci, nous a ramené à effectuer un redimensionnement des séries des silhouettes à une taille commune de 140 * 140 pixels. Enfin, Les moments de Zernike spatio-temporels sont calculés pour chaque volume de silhouettes.Dans les expériences, nous avons sélectionné le tiers des vidéos de chaque catégorie d'action pour l'étape d'apprentissage du classificateur LS-SVM et le reste sera utilisé pour les tests. Différents ordres de moment de Zernike ont été testés afin de choisir un ordre optimal permettant de résoudre efficacement le problème proposé. La matrice de confusion (a) montrent que l'ordre trois des moments de Zernike a donné des actions mal classés. Par exemple, l'action " jumping in place " présente un taux de bonne classification modeste (45%). La matrice (b) montre que l'ordre sept des moment de zernike a permis d'obtenir un bon taux de classification pour

	a1	a2	a3	a4	a5	a6	a7	a8	a9
a1	75	10	12	0	0	0	3	0	0
a2	27	55	0	0	0	0	8	10	0
a3	22	0	63	15	0	0	0	0	0
a4	20	3	0	55	4	3	4	5	6
a5	18	11	3	8	60	0	0	0	0
a6	13	12	13	0	17	45	0	0	0
a7	7	12	0	0	7	0	68	6	0
a8	21	0	9	0	0	0	9	61	0
a9	0	0	0	0	3	0	19	7	71

	a1	a2	a3	a4	a5	a6	a7	a8	a9
a1	100	0	0	0	0	0	0	0	0
a2	2	98	0	0	0	0	0	0	0
a3	7	13	80	0	0	0	0	0	0
a4	0	0	0	100	0	0	0	0	0
a5	10	0	0	0	83	7	0	0	0
a6	9	0	2	0	0	89	0	0	0
a7	0	0	0	0	2	0	96	0	2
a8	5	0	0	0	0	0	4	91	0
a9	0	0	0	0	0	0	10	7	83

FIG. 2 – *Matrice (a) ,Matrice (b)*

	a1	a2	a3	a4	a5	a6	a7	a8	a9
a1	88	11	0	0	0	1	0	0	0
a2	4	93	3	0	0	0	0	0	0
a3	4	8	66	0	11	11	0	0	0
a4	0	0	0	93	0	0	4	3	0
a5	0	0	7	0	91	2	0	0	0
a6	0	0	0	0	9	91	0	0	0
a7	0	0	0	0	0	0	90	10	0
a8	0	0	0	0	0	0	9	91	0
a9	0	0	3	0	3	0	0	0	94

	a1	a2	a3	a4	a5	a6	a7	a8	a9
a1	82	1	3	0	0	14	0	0	0
a2	2	35	51	0	10	2	0	0	0
a3	0	41	44	0	9	0	6	0	0
a4	0	0	0	96	0	1	0	1	2
a5	14	14	26	0	29	0	16	1	0
a6	0	0	0	13	0	85	0	1	1
a7	0	0	0	1	0	7	90	2	0
a8	0	1	0	1	0	0	44	52	2
a9	0	0	0	5	3	0	3	5	87

FIG. 3 – *Matrice (c) ,Matrice (d)*

presque toutes les actions. Pour prouver l'efficacité de ces résultats, nous avons comparés nos résultats à deux autres travaux (Dhillon et al (2009)) et (Kim et al (2007)). Les matrices de confusion (c) et (d) présentent les résultats des travaux de (Dhillon et al (2009)) et (Kim et al (2007)) sur la même base de données "Weizmann" . La comparaison de nos résultats expérimentaux avec les travaux de (Dhillon et al (2009)) et (Kim et al (2007)) prouve que notre approche a permis d'améliorer les résultats de classification pour la plupart des actions.

Dans les matrices, nous avons effectué les désignations suivantes : a1 : " walk ", a2 :" run ", a3 :"skip," a4 :"jack," a5 :"jump," a6 :"jump in place," a7 : "wave with one hand," a8 : "wave with two hands," and a9 :"bend")

FIG. 4 – *Exemple d'actions dans la base de données weizmann.*

4.2 Base de données d'actions KTH

Nous avons également mené des expériences sur la base de données d'actions KTH afin de confirmer les bonnes performances de notre approche. La base KTH contient six types d'actions humaines (boxer, agiter la main, claquer les mains, marcher, jogging et courir). Ces actions sont effectuées à plusieurs reprises par 25 personnes. Cette ensemble de données présente plus de complexité que l'ensemble des données Weizmann à cause des grandes variations des angles de vue, des échelles et d'apparences. Le même prétraitement appliqué aux vidéos de la base Weizmann à été également effectué sur la base de données KTH. La matrice de confusion (e) présente nos résultats de classification. La comparaison des résultats obtenus avec ceux d'autres méthodes comme Ning et al(2008) présenté par la matrice (f)montre que notre méthode présente un taux de classification meilleur. Cependant, nos résultats présentent des taux de classification comparable à ceux de la méthode de Kim et al (2007) présentés par la matrice de confusion (g).

	boxing	*handclappin*	*handwaving*	*jogging*	*runnig*	*walking*
boxing	91	7	2	0	0	0
handclappin	0	96	4	0	0	0
handwaving	3	2	80	0	0	15
jogging	4	0	0	85	10	1
runnig	0	0	0	0	85	15
walking	0	0	1	4	1	94

FIG. 5 – *Matrice (e)*

	boxing	*handclappin*	*handwaving*	*jogging*	*runnig*	*walking*
boxing	94	4	2	0	0	0
handclappin	1	95	3	0	0	1
handwaving	1	2	97	0	0	0
jogging	0	0	0	89	10	1
runnig	0	0	0	14	86	0
walking	0	0	0	5	1	94

FIG. 6 – *Matrice (f)*

	boxing	*handclappin*	*handwaving*	*jogging*	*runnig*	*walking*
boxing	98	2	0	0	0	0
handclappin	0	100	0	0	0	0
handwaving	1	2	97	0	0	0
jogging	0	0	0	90	10	0
runnig	0	0	0	12	88	0
walking	0	0	0	1	1	98

FIG. 7 – *Matrice (g)*

5 Conclusion

Nous avons proposé dans cet article une nouvelle approche pour la classification des actions dans les vidéos en se basant sur les moments de Zernike spatio-temporels. En effet, après la segmentation de la vidéo en entrée, une extraction des silhouettes a été effectuée sur les images des séquences vidéo. Ensuite, nous avons calculé les moments de Zernike spatio-temporels sur le volume des silhouettes obtenu. Enfin le classificateur LS-SVM a été utilisé pour classer les actions en se basant sur les moments de Zernike. La méthode proposée a été évaluée en réalisant des expériences sur les bases de données Weizmann et KTH. Les résultats expérimentaux montrent un bon taux de classification pour la plupart des actions. Dans les perspectives, nous allons étendre notre approche à la classification des vidéos contenant plusieurs personnes pratiquant des actions diverses. En plus, nous envisageons aussi de traiter l'effet d'occlusion sur la caractérisation des actions.

6 REFERENCES

Ali, A. and Aggarwal, J., "Segmentation and recognition of continuous human activity", in Procedure of Intelligent Workshop on Detection and Recognition of Events in Video, pp. 28-35, 2001.

Baccouche, M., Mamalet, F., Wolf, C. Garcia, C. andBaskurt, A., " Une approche neuronale pour la classification d'actions de sport par la prise en compte du contenu visuel et du mouvement dominant ", In Compression et Représentation des Signaux Audiovisuels (CORESA), pp.25-30, 2010.

Bobick, A. and Davis,J., "The recognition of human movement using temporal templates," IEEE Transaction on Pattern Analyses and Machine Intelligence,vol. 23, pp 257-267, 2001.

Blank,M., Gorelick, L., Shechtman, M., Irani, M. and Basri, R., " Actions as space-time shapes", Pattern Analysis and Machine Intelligence, IEEE Transactions ICCV,vol29, pp 2247-2253, 2005.

Cedras, C. and Shah, M., "Motion-based recognition : A survey," Image Vision Computer, vol. 13, pp. 129-155, 1995.

Dollar, P., Rabaud,V., Cottrell, G. and Belongie, S., "Behavior recognition via sparse spatio-temporal features" presented at the Int, Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance, 2005.

Dhillon, P. S., Nowozin, S., Lampert, C.H, Ť Combining appearance and motion for human action classification in videos Ş, Computer Vision and Pattern Recognition Workshops, 2009. CVPR Workshops 2009. IEEE Computer Society Conference on, 20-25 June 2009, pp : 22 - 29.

Dhillon, C. H., Nowozin, P.S., Lampert, S., " Combining appearance and motion for human action classification in videos", Computer Vision and Pattern Recognition Workshops, pp : 22 - 29, 2009.

Efros, A., Breg, C., Mori, G. and Malik, J. "Recognizing Action at a Distance", Computer Vision, Proceedings. Ninth IEEE International Conference,vol.2, pp 726-733. 2003.

Fihl, P., Corlin, R., Park, S., Moeslund, T., and Trivedi, M., Tracking of Individuals in Very Long Video Sequences. In Symposium on Visual Computing, Lake Tahoe, Nevada,USA, November 6-8 2006.

Gavrila, D., "The visual analysis of human movement : A survey" , Computer Vision Image Understand, vol. 73, pp. 82-98, 1999.

Gorelick,L., " al Actions as Space-Time Shapes" , Pattern Analysis and Machine Intelligence, IEEE Transactions, vol 29, pp. 2247-2253, 2007.

Guo, K., Ishwar, P.andKonrad, J., "Action Recognition in Video by Covariance Matching of Silhouette Tunnels", XXII Brazilian Symposium on Computer Graphics and Image Processing, pp. 299-306, 2009.

Green, R. and Guan, L., "Quantifying and recognizing human movement patterns from monocular video images," IEEE Transaction Circuits on System Video Technologys, vol. 14, pp. 179-190, 2004.

Kellokumpu, V., Pietikainen, M. and Heikkila, J., "Human activity recognition using sequences of postures" presented at the IAPR Conference on Machine Vision Applications, 2005.

Kim, T., Wong, S., Cipolla, R., "Tensor canonical correlation analysis for action classication", IEEE Conference on Computer Vision and Pattern Recognition (CVPR) pp. 1-3, 2007.

Kim,K., Chalidabhongse,T., Harwood,D. and Davis ,L., Real-time Foreground-Background Segmentation using CodebookModel. Real-time Imaging, vol 11,pp 167-256, 2005.

Liangliang ,C.,Ying,L., Zicheng ,L., Benjamin ,Y., and Thomas S.,ACTION DETECTION USING MULTIPLE SPATIAL-TEMPORAL INTEREST POINT FEATURES.

Ning, H., Han T.,Walther D., Liu M., Huang T.," Hierarchical space-time model enabling efficient search for human actions". IEEE Transactions on Circuits and Systems for Video Technology, in press,(2008).

Rao ,C.and Shah,M., "View-invariance in action recognition", in Proc IEEE Conf. Computer Vision and Pattern Recognition, vol. 2, pp.316-321,2001.

Song,Y., Goncalves,L. and Perona,P., "Unsupervised learning of human motion," IEEE Transaction on Pattern Analyses and Machine Intelligence, vol. 25,pp. 814-827, 2003.

Sminchisescu,C., Kanaujia,A., Li,Z., and Metaxas,D., "Conditional models for contextual human motion recognition," in Proc. Int. Conf.Computer Vision, vol. 2, pp 1808-1815, 2005.

Shutler ,J. D. and Nixon,M. S."Zernike velocity moments for the description and recognition of moving shapes".Proc. British Machine Vision Conference (BMVC01), 2 :pp. 705-714, 2001.

Suykens ,J.A.K. and Vandewalle,J., "Least Squares Support Vector Machines" ,Neural Processing Letters 3,Vol 9, pp.293-300, 2002.

Teague,M. R."Image analysis via the general theory of moments", Journal of the Optical Society of America, 70(8) :pp. 920-930, 1979.

Van Gestel, J., & al., "Benchmarking Least Squares Support Vector Machine Classifiers", Technical report, Internal Report 00-37, ESAT-SISTA, K.U.Leuven (Leuven, Belgium) 2000.

Weinland,D., Ronfard,R. and Boyer,E., "Motion history volumes for free viewpoint action recognition" presented at the IEEE Workshop Modeling People and Human Interaction, pp 87-89, 2005.

Wu,X., "Templated-based Action Recognition : Classifying Hockey Players Movement", Master's thesis, The University of British Columbia, 2005.

Yacoob, Y. and Black,M., "Parameterized modeling and recognition of activities," Computer Vision on Image Understand", vol. 73,p. 232-247, 1999.

Zelnik-Manor, L. and Irani,M., "Event-Based Analysis of Video", Computer Vision and Pattern Recognition,Computer Vision and Pattern Recognition,Proceedings of the 2001 IEEE Computer Society Conference vol 2, pp.123-130,2001.

Summary

Action recognition in video and still image is one of the most challenging research topics in pattern recognition and computer vision. This paper proposes a new method for video action classification based on space-time Zernike moments. This moments aim to capturing both structural and temporal information of a time varying sequence. The originality of this approach consists to represent actions in video sequences by a three-dimension shape obtained from different silhouettes in the space-time volume. In fact, the given video is segmented in space-time volume. Then, silhouettes are extracted from obtained images of the video sequences volumes and a space-time Zernike moments are computed for silhouettes volumes. Finally, least square version of SVM (LS-SVM) classifier with extracted features is used to classify actions in videos. To evaluate the proposed approach, it was applied on a benchmark

human action dataset Weizmann and KTH. The experimentations and evaluations show efficient results in terms of action characterizations and classification. Furthermore, it presents several advantages such as simplicity and respect of silhouette movement progress in the video guaranteed by space-time Zernike moments.