Reconnaissance de gestes : approches 2D & 3D

Maher Mkhinini et Patrick Horain
Institut Mines-Télécom/Télécom SudParis
Département Électronique et Physique, 9 rue Charles Fourier, 91011 Evry, France
Email : {prénom.nom}@telecom-sudparis.eu

Résumé

Dans cet article nous comparons deux approches pour la reconnaissance de gestes en 2D et en 3D. La première exploite les points d'intérêt spatio-temporels où chaque séquence est décrite par des histogrammes de flux optique locaux. La reconnaissance est effectuée par une machine à vecteur support (SVM). La deuxième approche consiste à utiliser la trajectoire 3D de la main acquise à l'aide d'un capteur Kinect et la reconnaissance est effectuée par déformation temporelle dynamique (DTW). Nous comparons les deux approches en étudiant leurs limitations, l'efficacité de chacune en termes de taux de reconnaissance et de temps de calcul.

Mots clés

Interaction homme-robot, reconnaissance de gestes, Kinect, SVM, DTW.

1. Introduction

Les évolutions des capteurs et l'élargissement des domaines d'utilisation ont fait croitre l'importance de la reconnaissance de gestes par vision artificielle. En raison du grand nombre de degrés de liberté du corps humain, l'interprétation des gestes par vision est difficile car leur apparence dépend de la direction d'observation, de la trajectoire des parties significatives du corps, en plus des difficultés classiques en vision, notamment la variation de luminosité, contraste, variation d'échelle.
Plusieurs approches ont montré leur efficacité pour la reconnaissance et la classification des mouvements humains (gestes, actions...). Certaines procèdent par détection et suivi des trajectoires du corps entier ou des parties du corps, d'autres par l'analyse géométrique et temporelle des vidéos et l'extraction des caractéristiques d'apparence dans les images.

Bobick et al. [1] proposent de reconnaître des activités sur une mesure de correspondance entre signatures gestuelles, *Motion History Image* (MHI) et *Motion Energy Image* (MEI). Le MEI représente le geste ou l'action à partir de la variation temporelle des pixels dans les régions de mouvement et l'intensité spatio-temporelle de mouvement. Alors que le MHI représente la forme de la région de mouvement dans l'image. Afin de caractériser une séquence, ils calculent les vecteurs de moment des formes obtenues dans le MHI et le MEI. Pour classer les gestes, ils mesurent la distance de Mahalanobis entre les modèles de référence et le modèle calculé sur les images acquises. L'avantage de cette méthode est la prise en compte des évolutions spatiales et temporelles des gestes. Toutefois, la comparaison des signatures gestuelles qui sont représentées dans un espace de grande dimension nécessite une grande puissance de calcul. Il existe d'autres approches qui sont basées sur l'extraction des caractéristiques sur des régions locaux dans l'image [2] [3]. Deux étapes sont généralement nécessaires pour la reconnaissance : apprentissage de gestes à partir d'une base de vidéos et la classification où les descripteurs extraits sont comparés aux descripteurs d'apprentissage.
Kaaniche et al. [4] ont développé une approche de reconnaissance de gestes en utilisant un système multi-caméras. Ils détectent les points d'intérêts dans la scène afin d'extraire les régions caractéristiques et les décrivent par des histogrammes de gradient orienté *HOG* [5]. Ensuite, ils effectuent un suivi temporel de ces descripteurs et classent les gestes en utilisant l'algorithme *K-means* à partir d'une base de gestes d'apprentissage. La reconnaissance est réalisée par un algorithme de plus proches voisins (*K-nearest neighbors*). Le caractère dynamique des gestes est décrit en temps réel par le suivi des points détectés par HOG dans la séquence. Cependant, l'utilisation des HOG comme caractéristique rend cette approche sensible aux changements d'échelle et surtout aux rotations de la personne qui engendrent un grand changement

d'apparence.
Dans cet article, nous proposons une évaluation de deux approches de reconnaissance de gestes qui généralisent les deux grands axes de recherche dans ce domaine. La première approche consiste à utiliser des points caractéristiques spatio-temporelle (STIP) [6] afin de décrire localement les mouvements par des histogrammes de flux optique qui sont ensuite classés par une machine à vecteurs support (*SVM*). La deuxième approche est basée sur la classification des trajectoires 3D de la main. Chaque geste effectué est représenté par les coordonnées 3D de la main en fonction du temps. La classification est effectuée par le calcul de la déformation temporelle dynamique (DTW). La section suivante de cet article est consacrée à la description de l'approche 2D. La section 3 décrit l'approche 3D avec la déformation temporelle dynamique pour la classification des gestes. Dans la section 4, nous décrivons le corpus vidéo utilisé pour tester les deux approches. La dernière partie présente les résultats obtenus.

2. Points d'intérêt spatio-temporels

Les points d'intérêts spatio-temporels (STIP) sont de plus en plus utilisés dans la reconnaissance des actions et des mouvements humains. Ivan Laptev [6] a proposé une extension de la détection des points d'intérêts spatiaux de Harris [7] et Forstner [8] pour une détection spatio-temporelle des régions locales où il existe une forte variation conjointe spatio-temporelle. En vision, les points d'intérêts correspondent à une forte variation spatiale locale(les coins, les intersections, des textures...). Ces variations sont généralement des contours intenses dans l'image ou des coins. Harris et al. [7] proposent de détecter les régions de variation significative dans l'image en déterminant les points qui maximisent la fonction de coin H^{sp} qui correspond à la variation de second ordre de la fonction de courbure. Pour chaque point $p(x, y)$ dans l'espace d'une image H^{sp} est définie par :

$$H^{sp} = \det(\mu^{sp}) - k \times trace^2(\mu^{sp}) \qquad (1)$$

où μ^{sp} est la matrice de moment de deuxième degré en $p(x, y)$ et $k \in R$.

Laptev [6] propose une extension à un détecteur de points d'intérêt spatio-temporels (STIP) en maximisant la fonction H définie pour chaque point $p(x, y, \tau)$ dans le domaine spatio-temporel par :

$$H = \det(\mu) - k \times trace^3(\mu) \qquad (2)$$

où μ est la matrice de moment spatio-temporel en $p(x, y, \tau)$ et $k \in R$.

Extraction des caractéristiques. Afin de décrire le geste dans une séquence donnée nous avons calculé des histogrammes locaux de flux optique (HoF) sur les régions de voisinage des STIP détectés [9]. Chaque bloc de pixels (spatio-temporel) est décrit par un histogramme sur une grille de taille $\eta_x \times \eta_y \times \eta_\tau$ avec $\eta_x = \eta_y = 3$ et $\eta_\tau = 2$. Pour chaque bloc on calcule un histogramme normalisé de flux optique de 5 accumulateurs, ce qui donne finalement un vecteur de dimension 90. Chaque séquence génère un nombre d'histogrammes égal au nombre de point d'intérêt détectés. Nous avons utilisé la technique de la création de vocabulaires visuels (*Bag-of-Features, BoF*) à partir des histogrammes HoF calculés et l'apprentissage d'une machine SVM multi-classes non linéaire.

BoF pour la reconnaissance de gestes. La technique de vocabulaire visuel *(Bag of Features) (BoF)* est très connue dans la reconnaissance d'objet et l'indexation des images [10], cette méthode est de plus en plus utilisée dans le domaine de classification des actions et des gestes. Elle permet de créer un modèle génératif pour la représentation et la description locale des classes d'objet. Cette méthode nécessite trois étapes, la création du vocabulaire visuel, la quantification des données d'apprentissage et la création des histogrammes de fréquence d'apparence.
Nous avons utilisé la méthode *K-means* pour la quantification des histogrammes HoF avec K=1000 clusters. *K-means* est appliqué sur un tiers des données d'apprentissage sélectionnées aléatoirement. Cette étape fournit un dictionnaire d'histogramme qui représente l'ensemble des classes de geste de notre corpus qu'on va appeler un vocabulaire visuel.
Ensuite, nous attribuons chaque vecteur des données d'apprentissage au plus proche vecteur de notre vocabulaire visuel en utilisant la distance euclidienne. Puisque chaque séquence est décrite par un ensemble d'histogramme HoF, un histogramme de fréquence d'apparence de ces HoF dans le vocabulaire visuel est calculé.

SVM multi-classes non linéaire. Nous utilisons pour la classification une machine à vecteurs support multi-classe non-linéaire [11] [12]. Nous avons utilisé la librairie LIBSVM [13] avec un noyau gaussien (*radial basis function RBF*) qui pour tout $x = x_1, \ldots, x_N$ et $x' = x'_1, \ldots, x'_N$ est défini par :

$$K(x, x') = \exp(\gamma \times |x - x'|^2) \qquad (3)$$

Avec γ est un paramètre du noyau qu'on peut le déterminer pratiquement en utilisant la validation croisée (*cross validation*). Cependant, γ peut être exprimé en fonction de la variance des données d'apprentis-

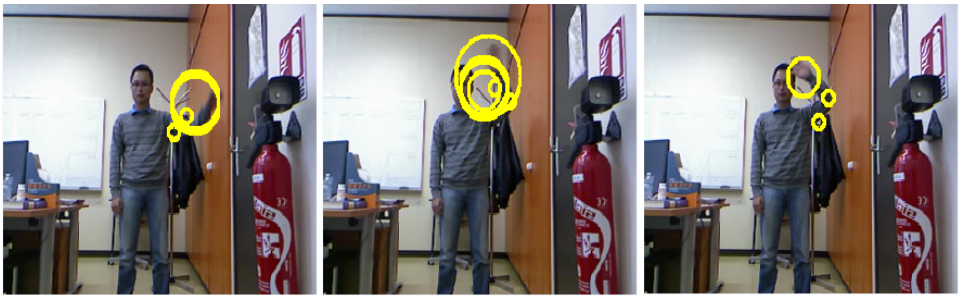

FIGURE *1: Exemple de détection des STIP sur la séquence de geste « Salut ».*

sage [11] ($\gamma = \frac{1}{2\sigma^2}$). Dans notre cas nous avons obtenu une valeur de $\gamma \simeq 0.1$ sur la base d'apprentissage décrite dans la section 4.

3. Reconnaissance de gestes par le suivi 3D de la main

L'analyse et l'interprétation des mouvements humains peuvent aussi être effectuées par des primitives 3D (positions, angles articulatoires...) qui nécessitent un suivi 3D du corps entier ou de quelque partie du corps selon les mouvements à interpréter [14] [15].
Nous proposons de classer des gestes dynamiques à partir de la trajectoire 3D de la main. La reconnaissance est effectuée par la mesure de la déformation temporelle dynamique [16].

Déformation temporelle dynamique. Le DTW est une technique de mesure de distance entre deux séquences qui ont une variation temporelle (exécutées à une vitesse différente par exemple). Elle a été utilisé dans la reconnaissance de la parole [17] et a été appliqué dans la reconnaissance des gestes [18] [19].

Soient deux séries numériques temporelles U et V, de longueur respective n et m, avec :

$$U = u_1, u_2, ..., u_i, ..., u_n \tag{4}$$

$$V = v_1, v_2, ..., v_j, ..., v_m \tag{5}$$

Pour aligner les deux séquences par la technique DTW nous calculons la matrice $(n - sur - m)$ avec le (i^{eme}, j^{eme}) élément de cette matrice contient la distance $d(u_i, v_j)$ entre les deux points u_i et v_j. Dans notre cas nous utilisons la distance euclidienne. Le chemin de déformation W correspond aux éléments de la matrice en la parcourant du premier point de la séquence $(1, 1)$ en atteignant le dernier point (n, m).

$$W = w_1, w_2, ..., w_k, ..., w_K \tag{6}$$

avec, $max(m, n) \leq K < m + n - 1$

Il existe plusieurs chemins de déformation possible, le DTW correspond au chemin qui minimise le cumule des distances locales.

$$DTW(U, V) = \min_{k \leq K} \frac{\sqrt{\sum_{k=1}^{K} w_k}}{K} \tag{7}$$

Reconnaissance de gestes par DTW. En reconnaissance de gestes, cette technique consiste à comparer une séquence de test à des séquences de référence. Chaque séquence est représentée par les coordonnées 3D de la main. Kinect fournit les coordonnées 3D du corps par rapport à un repère qui est situé au centre du capteur. L'utilisation de ces coordonnées brutes rend notre système dépendant de la position de l'acteur par rapport au capteur. Pour cela, nous avons utilisé les coordonnées de la main par rapport au repère du corps. L'origine de ce repère est le bas du buste.

4. Corpus vidéo

Le vocabulaire gestuel défini dans notre étude est composé de quatre gestes liés à des commandes vocales. Voici la description des gestes :

- « Bonjour » : mouvement de l'avant bras de droite à gauche dans un plan frontal.
- « Stop » : main ouverte dans un plan frontal avec bras tendu.
- « Va-t-en » : mouvement de l'avant bras d'arrière en avant dans un plan sagittal.
- « Approche » : mouvement de l'avant bras d'avant en arrière dans un plan sagittal.

Chaque geste est interprété dix fois par trois personnes différentes. Les séquences de gestes sont acquises par le capteur Kinect afin d'extraire, en plus des séquences d'image, les trajectoires 3D de la main droite et le centre du corps donnés par le suivi en temps réel du squelette. Le corpus contient 120 séquences ce qui correspond à environ 20 minutes d'enregistrement. En total, nous avons 30 séquences par classe de geste. Dans notre évaluation de la méthode de reconnaissance

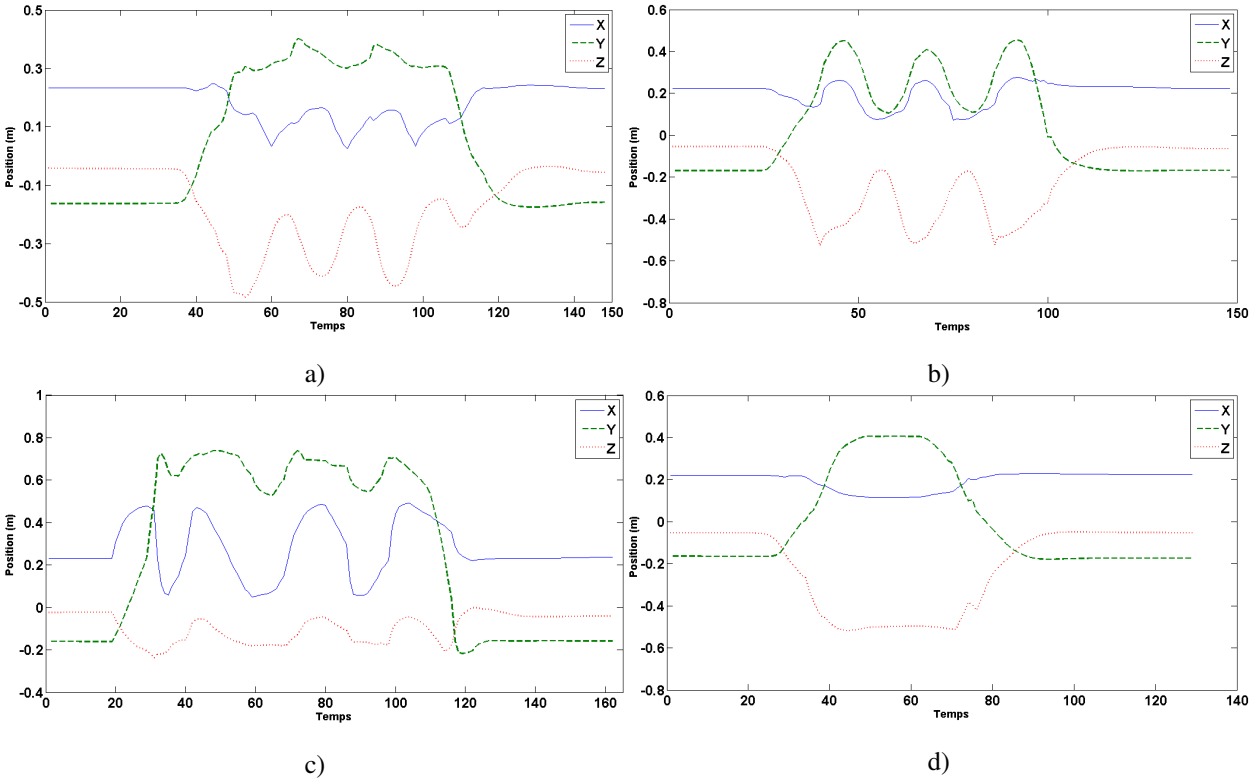

FIGURE 2: *Exemple des coordonnées X, Y, Z de la main pour chaque type de gestes, a)« Approche », b)« Va-t-en », c)« Bonjour », d)« Stop ».*

	Bonjour	Approche	Va-t-en	Stop
Bonjour	**100%**	-	-	-
Approche	-	**100%**	-	-
Va-t-en	-	-	**100%**	-
Stop	-	-	-	**100%**

Tableau 1: Matrice de confusion en utilisant les STIP

	Bonjour	Approche	Va-t-en	Stop
Bonjour	**92%**	8%	-	-
Approche	-	**100%**	-	-
Va-t-en	-	-	**100%**	-
Stop	-	-	-	**100%**

Tableau 2: Matrice de confusion en utilisant le DTW

en utilisant les STIP nous avons utilisé six séquences par personne par classe de geste dans l'apprentissage et quatre séquences par personne par classe de geste pour le test. En total nous avons 72 séquences d'apprentissage et 48 séquences de test. Pour l'évaluation de la technique de DTW, nous avons sélectionné aléatoirement trois séquences de référence. Les sept séquences restantes sont utilisées dans la phase de test.

5. Résultats expérimentaux

Nous avons comparé les deux approches 2D et 3D sur notre corpus. Les résultats obtenus sont représentés dans les deux matrices de confusion ci-dessus.

La matrice de confusion de l'évaluation de la reconnaissance de gestes par les STIP et l'apprentissage d'une machine SVM multi-classes non linéaire (Tableau 1) montre un taux de reconnaissance égal à 100% dans toutes les classes de gestes. Les séquences de test et d'apprentissage sont choisies aléatoirement dans notre corpus. Aucun prétraitement des séquences n'a été effectué. Ce résultat montre l'efficacité des caractéristiques locales extraites en tenant compte de l'évolution temporelle des mouvements. En effet, l'aspect dynamique des gestes nécessite la prise en compte de l'évolution spatiale et temporelle des mouvements et pas seulement des caractéristiques géométriques dans l'image. Cependant, cette méthode repose entièrement sur la détection des STIP qui est invariante à la rotation de l'image. Un autre inconvénient est le temps de calcul, cette méthode nécessite 2 secondes pour détecter les STIP sur une séquence d'images de dimension 640x480 et de longueur 7 seconde (210 images).
Dans le cas de la reconnaissance sur les trajectoires

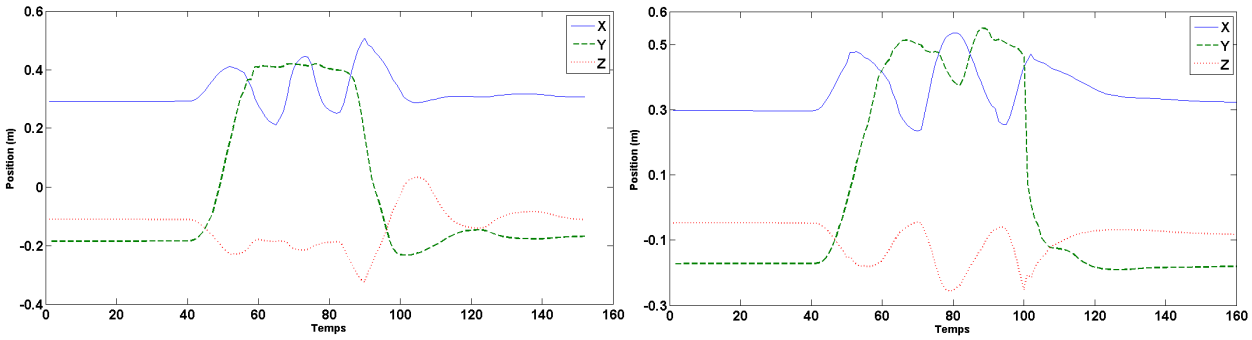

FIGURE *3: Comparaison entre une séquence de test (a) et de référence (b) du geste « Bonjour », réalisées par la même personne.*

3D de la main, la déformation temporelle dynamique donne des résultats satisfaisants (Tableau 2) malgré une confusion entre le geste « Bonjour » et « Approche » qu'on ne trouve pas dans l'approche 2D utilisant les STIP. L'une des causes de ces résultats est l'imprécision d'acquisition du capteur Kinect sachant que les données acquises sont relativement bruitées à cause des imprécisions de suivi des articulations. La différence d'exécution des gestes est aussi une des causes des erreurs de classification données par cette méthode comme le montre la figure 3. L'avantage de la méthode 3D par rapport à celle en 2D est le temps de calcul, elle nous permet de classer une séquence en temps réel. Mais les limitations sont principalement liées au capteur kinect à cause du bruit et de la distance maximale de l'acteur par rapport au capteur qui est limitée à 4 mètres pour avoir une précision suffisante dans l'image de profondeur pour le suivi de squelette.

6. Conclusion

Nous avons présenté deux approches différentes de reconnaissance de gestes dynamiques. L'approche 2D consiste à décrire les points d'intérêt spatio-temporels détectés sur toute la séquence de geste par des histogrammes de flux optique. Chaque séquence est ensuite décrite par un histogramme de fréquence en utilisant la technique de vocabulaire visuel (Bof). L'apprentissage est effectuée par une machine SVM multi-classe non linéaire. La deuxième approche est la mesure de déformation dynamique temporelle entre les séquences de geste. Le test de cette méthode sur la trajectoire 3D de la main donne des résultats satisfaisants, légèrement moins bon que les résultats obtenus dans l'approche 2D. Toutefois, l'approche 2D ne permet pas d'avoir une classification en temps réel. Nous poursuivons actuellement notre étude sur des gestes plus complexes.

Remerciement

Ce travail a été réalisé dans le cadre du projet Juliette financé par le FEDER (Fonds Européen de Développement Régional) sous la convention CRIF 10012367/R.

Références

[1] A. Bobick and J. Davis, "The recognition of human movement using temporal templates," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 23, no. 3, pp. 257–267, 2001.

[2] V. Pavlovic, R. Sharma, and T. Huang, "Visual interpretation of hand gestures for human-computer interaction : a review," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 19, no. 7, pp. 677–695, 1997.

[3] J. Alon, V. Athitsos, Q. Yuan, and S. Sclaroff, "A unified framework for gesture recognition and spatio-temporal gesture segmentation," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 31, no. 9, pp. 1685–1699, 2009.

[4] M. B. Kaaniche and F. Bremond, "Tracking hog descriptors for gesture recognition," in *Proceedings of the 2009 Sixth IEEE International Conference on Advanced Video and Signal Based Surveillance*, ser. AVSS '09. Washington, DC, USA : IEEE Computer Society, 2009, pp. 140–145.

[5] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, vol. 1, 2005, pp. 886–893.

[6] I. Laptev, "On space-time interest points," *Int. J. Comput. Vision*, vol. 64, no. 2-3, pp. 107–123, Sep. 2005.

[7] C. H. and M. S., "A combined corner and edge detector," in *In Proc. of Fourth Alvey Vision Conference*, 1988, pp. 147–151.

[8] W. Förstner and E. Gülch, *A fast operator for detection and precise location of distinct points corners and centres of circular features*. ISPRS, 1987, pp. 281–305.

[9] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld, "Learning realistic human actions from movies," in *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, 2008, pp. 1–8.

[10] S. Agarwal and D. Roth, "Learning a sparse representation for object detection," in *Proceedings of the 7th European Conference on Computer Vision-Part IV*, ser. ECCV '02. London, UK, UK : Springer-Verlag, 2002, pp. 113–130.

[11] J. Zang, M. Marsza, S. Lazebnik, and C. Schmid, "Local features and kernels for classification of texture and object categories : A comprehensive study," *Int. J. Comput. Vision*, vol. 73, no. 2, pp. 213–238, Jun. 2007.

[12] C. Schuldt, I. Laptev, and B. Caputo, "Recognizing human actions : a local svm approach," in *Pattern Recognition, 2004. ICPR 2004. Proceedings of the 17th International Conference on*, vol. 3, 2004, pp. 32–36.

[13] C.-C. Chang and C.-J. Lin, "Libsvm : A library for support vector machines," *ACM Transactions on Intelligent Systems and Technology*, vol. 2, pp. 1–27, 2011. [Online]. Available : http ://www.csie.ntu.edu.tw/ cjlin/libsvm

[14] H. Wang, M. M. Ullah, A. Klaser, I. Laptev, and C. Schmid, "Evaluation of local spatio-temporal features for action recognition," in *Proc. BMVC*, 2009, pp. 124.1–124.11.

[15] A. Just and S. Marcel, "A comparative study of two state-of-the-art sequence processing techniques for hand gesture recognition," *Comput. Vis. Image Underst.*, vol. 113, no. 4, pp. 532–543, Apr. 2009.

[16] C. S. Myers and R. Rabiner, L., "A comparative study of several dynamic time-warping algorithms for connected word recognition," *The Bell System Technical Journal*, pp. 1389–1409, Sep. 1981.

[17] H. Sakoe and S. Chiba, "Dynamic programming algorithm optimization for spoken word recognition," *Acoustics, Speech and Signal Processing, IEEE Transactions on*, vol. 26, no. 1, pp. 43–49, 1978.

[18] D. Gavrila and L. Davis, "3-d model-based tracking of humans in action : a multi-view approach," in *Computer Vision and Pattern Recognition, 1996. Proceedings CVPR '96, 1996 IEEE Computer Society Conference on*, 1996, pp. 73–80.

[19] J. Lichtenauer, E. Hendriks, and M. J. T. Reinders, "Sign language recognition by combining statistical dtw and independent classification," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 30, no. 11, pp. 2040–2046, 2008.