Programa de Doctorat:

AUTOMÀTICA, ROBÒTICA I VISIÓ

Tesi Doctoral

A template based approach for human action recognition

Josep Maria Carmona Leyva

Director: Dr. Joan Climent Vilaró

Barcelona, abril de 2018

A la meva dona i fills pel seu recolzament i col·laboració, als meus pares per mostrar-me el valor de l'esforç.

Investigar és veure el que tothom ha vist, i pensar el que ningú més ha pensat.

Albert Szent (1893-1986)

Qualsevol tecnologia suficientment avançada és indistingible de la màgia.

Arthur C. Clarke (1917-2008)

Abstract

Visual analysis of human movements concerns the understanding of human activities from image sequences. The goal of the action/gesture recognition is to recognize the label that corresponds to an action or gesture made by a human in a sequence of images.

To solve this problem, the researchers have proposed solutions that range from object recognition techniques, to speech recognition techniques, face recognition or brain function.

The techniques presented in this thesis, are related to a set of techniques that condense a video sequence into a template that retain important information to action/gestures classification applying standard object recognition techniques.

In a first stage of this thesis, we have proposed a view-based temporal template approach for action/gesture representation from tensors. The templates are computed from three different projections considering a video sequence as a third-order tensor. We compute each projection from the fibers of the tensor using a combination of simple functions. We have studied which function and feature extractor/descriptor is the most suitable to project the template from the tensor. We have tested five different simple functions used to project the fibers, namely, supremum, mean, standard deviation, skewness and kurtosis using public datasets. We have also studied the performance obtained applying four feature extractors/descriptors like PHOW, LIOP, HOG and SMFs.

Using more complex datasets, we have assessed the most suitable feature representation for our templates (Bag Of Words or Fisher Vectors) and the complementarity among the features computed from each simple function (*Max*, *Mean*, *Standard Deviation*, *Kurtosis* y *Skewness*). Finally, we have studied the complementarity with a successful technique like Improved Dense Trajectories.

The experiments have shown that Standard Deviation function and PHOW extractor/descriptor are the most suitable for our templates. The results have shown also that our 3 projection templates overcome most state-of-the-art techniques in more complex datasets when we combine the templates with Fisher Vector representation. The features extracted by each simple function are complementary among them and that added to HOG, HOF and MBH improves the performance of IDTs.

Derived from this thesis, we have also presented another view-based temporal template approach for action recognition obtained from a Radon transform projection and that allows the temporal segmentation of human actions in real time. First, we propose a generalization of the *R* transform that it is useful to adapt the transform to the problem to be solve. We have studied the

performance in three functions, namely, *Max, Mean and Standard Deviation* for pre-segmented human action recognition using a public dataset, and we have compared the results against traditional *R* transform. The results have shown that *Max* function obtains the best performance when it is applied on Radon transform and that our technique overcomes many state-of-the-art techniques in action recognition.

In a second stage, we have modified the classifier to adapt it to temporal segmentation of human actions. To assess the performance, we have merged Weizman and Hollywood actions datasets and we have measured the performance of the method to identify isolated actions. The experiments have shown that our technique overcomes the state-of-the-art techniques in Weizman dataset in no pre-segmented human actions.

Resumen

El análisis visual de movimientos humanos hace referencia al entendimiento de la actividad humana en secuencias de video. El objetivo del reconocimiento de acciones/gestos en ámbito de la Visión por Computador, es identificar el nombre que corresponde a una acción o gesto realizado en una secuencia de imágenes.

Para dar solución a este problema, los investigadores han propuesto soluciones que van desde la aplicación de técnicas que derivan del reconocimiento de objetos, del reconocimiento del habla, del reconocimiento facial o del funcionamiento del cerebro.

Las técnicas presentadas en esta tesis, están relacionadas con un conjunto de técnicas que intentan condensar una secuencia de video en unas *templates* que retienen información importante de cara a la discriminación entre acciones/gestos aplicando técnicas estándar de reconocimiento de objetos.

En la primera parte de esta tesis, hemos propuesto una aproximación basada en *template* para la representación de acciones/gestos a partir de tensores. Nuestras *templates* se calculan desde tres proyecciones diferentes considerando una secuencia de vídeo como un tensor de tercer orden. Calculamos cada proyección desde las fibras del tensor de tercer orden utilizando funciones simples. Hemos hecho un estudio exhaustivo para encontrar qué función debe ser utilizada para proyectar el *template* desde el tensor, y qué extractor/descriptor es el más adecuado. Utilizando *datasets* públicos simples, hemos testeado cinco funciones diferentes simples para proyectar las fibras, llamadas, *Max*, *Mean*, *Standard Deviation*, *Kurtosis* y *Skewness*. Hemos estudiado también el rendimiento obtenido aplicando a nuestras templates, cuatro técnicas de extracción/descripción de características del estado del arte como PHOW, LIOP, HOG y SMFs.

Utilizando *datasets* más complejos, hemos estudiado cuál es la mejor representación de las características extraídas de las *templates* (*Bag Of Words* o *Fisher Vectores*), y la complementariedad entre las características extraídas con cada una de las cinco funciones (*Max*, *Mean*, *Standard Deviation*, *Kurtosis* y *Skewness*) y la complementariedad de estas con una exitosa técnica como *Improved Dense Trajectories*.

Los experimentos han demostrado que la desviación estándar es la mejor función para proyectar las fibras en las *templates*, y que PHOW obtiene el mejor rendimiento como detector/descriptor en las *templates* obtenidas. Los *datasets* más complejos han mostrado que la mejor representación para las características extraídas de las *templates* es Fisher Vectores, que existe complementariedad entre las características extraídas con cada una de las funciones y que la

fusión de estas características con Improved Dense Trajectories, hace que este último mejore su rendimiento.

Derivado de los trabajos de esta tesis, también presentamos otra aproximación basada en *template* por el reconocimiento de acciones/gestos que se obtiene de una proyección derivada de la transformada de Radon y que permite la segmentación temporal de acciones en tiempo real. Primero hemos planteado una generalización de la transformada *R* que permite adaptar la transformada al problema a resolver mediante la función de proyección. Hemos estudiado su rendimiento para las funciones *Max*, *Mean* y *Standard Deviation* en reconocimiento de acciones pre-segmentadas sobre un *dataset* público y comparado los resultados con la transformada *R*. Los resultados han mostrado que la función *Max* obtiene el mejor resultado cuando se aplica sobre la transformada de Radon y que nuestra técnica supera a muchos métodos del estado del arte en reconocimiento de acciones.

En una segunda fase, hemos introducido una modificación en la etapa de clasificación de nuestra técnica para permitir segmentar acciones temporalmente. Para evaluar su rendimiento, hemos concatenado acciones de los *datasets* Weizmann y Hollywood y medido la capacidad de la técnica para identificar cada una de las acciones individuales. Los experimentos han demostrado que nuestra técnica rinde mejor en la segmentación de acciones del Weizmann *dataset* que las técnicas del estado del arte.

Resum

L'anàlisi visual de moviments humans fa referència al enteniment d'activitat humana en seqüències de vídeo. L'objectiu del reconeixement d'accions/gestos en l'àmbit de la Visió per Computador, és identificar el nom que correspon a una acció o gest realitzat en una seqüència d'imatges.

Per donar solució a aquest problema, els investigadors han proposat solucions que van des de l'aplicació de tècniques que deriven del reconeixement d'objectes, del reconeixement de la parla, del reconeixement facial o del funcionament del cervell.

Les tècniques presentades en aquesta tesi, estan relacionades amb un conjunt de tècniques que intenten condensar una seqüència de vídeo en uns *templates* que retinguin informació important de cara a la discriminació entre accions/gestos aplicant tècniques estàndards de reconeixement d'objectes.

A la primera part d'aquesta tesi, hem proposat una aproximació basada en *template* per la representació d'accions/gestos a partir de tensors. Les nostres *templates* es calculen des de tres projeccions diferents considerant una seqüència de vídeo com un tensor de tercer ordre. Calculem cada projecció des de les fibres del tensor de tercer ordre utilitzant funcions simples. Hem fet un estudi exhaustiu per trobar quina funció ha de ser utilitzada per projectar el *template* des del tensor, i quin extractor/descriptor és el més adequat. Utilitzant *datasets* públics simples, hem testejat cinc funcions diferents simples per projectar les fibres, anomenades, *Max*, *Mean*, *Standard Deviation*, *Kurtosi* i *Skewness*. Hem estudiat també el rendiment obtingut aplicant a les nostres *templates*, quatre tècniques d'extracció/descripció de característiques de l'estat de l'art com PHOW, LIOP, HOG i SMFs.

Utilitzant *datasets* més complexes, hem estudiat quina és la millor representació de les característiques extretes de les *templates* (*Bag Of Words* o *Fisher Vectors*) i la complementarietat entre les característiques extretes amb cada una de les cinc funcions (*Max*, *Mean*, *Standard Deviation*, *Kurtosi* i *Skewness*) i la complementarietat d'aquestes amb una exitosa tècnica com *Improved Dense Trajectories*.

Els experiments han demostrat que la desviació estàndard és la millor funció per projectar les fibres en les *templates*, i que PHOW obté el millor rendiment com a detector/descriptor en les *templates* obtingudes. Els *datasets* més complexes han mostrat que la millor representació per a les característiques extretes de les *templates* és amb *Fisher Vectors*, que existeix complementarietat entre les característiques extretes amb cada una de les funcions i que la fusió

d'aquestes característiques amb *Improved Dense Trajectories*, fa que aquest últim millori el seu rendiment.

Derivat dels treballs d'aquesta tesi, també presentem una altre aproximació basada en *template* pel reconeixement d'accions/gestos que s'obté d'una projecció derivada de la transformada de Radon i que permet la segmentació temporal d'accions en temps real. Primer hem plantejat una generalització de la transformada *R* que permet adaptar la transformada al problema a resoldre mitjançant la funció de projecció. Hem estudiat el seu rendiment per a les funcions *Max*, *Mean* i *Standard Deviation* en reconeixement d'accions pre-segmentades sobre un *dataset* públic i comparat els resultats amb la transformada *R*. Els resultats han mostrat que la funció *Max* obté el millor resultat quan s'aplica sobre la transformada de Radon i que la nostra tècnica supera a molts mètodes de l'estat de l'art en reconeixement d'accions.

A una segona fase, hem introduït una modificació a la etapa de classificació de la nostra tècnica per permetre segmentar accions temporalment. Per avaluar el seu rendiment, hem concatenat accions dels *datasets* Weizmann i Hollywood i mesurat la capacitat de la tècnica per identificar cadascuna de les accions individuals. Els experiments han demostrat que la nostra tècnica rendeix millor en la segmentació de les accions del *dataset* Weizmann que les tècniques de l'estat de l'art.

Contingut

Llista de taules

Llista de figures

1. Introducció

L'anàlisi visual de moviments humans fa referència al enteniment d'activitat humana en seqüències de vídeo. La classificació i localització automàtica d'accions humanes o gestos és útil per múltiples aplicacions com video-vigilància, interacció humana amb computadors, indexació de vídeo o anàlisi per biometria. Tècnicament, una acció és una seqüència de moviment generada per un humà durant la realització d'una tasca. Així una acció o gest es una entitat de quatre dimensions que pot ser descomposta en les seves parts espacial i temporal. Els humans donem nom a les accions i gestos que realitzem de forma que generalment la majoria de les persones poden saber i realitzar l'acció a partir del seu nom. L'objectiu del reconeixement d'accions/gestos és identificar el nom que correspon a una acció o gest realitzat. Idealment, aquest reconeixement ha de ser invariant a la velocitat de realització de l'acció, distància de l'observador, etc.

Les tècniques plantejades pels investigadors per intentar solucionar aquest tipus de problemes, generalment consten de les següents etapes [3]:

- Extracció de característiques: és una de les tasques principals del reconeixement d'accions. Consisteix en localitzar regions de les imatges que aportin informació de moviment, forma o altres tipus, que permetin discriminar entre les diferents accions.
- Aprenentatge i classificació: Tracta de les tasques d'aprenentatge de models estadístics que permetran més tard classificar correctament accions realitzades per persones o en entorns no vistos prèviament a la fase d'aprenentatge. Aquesta fase ha de ser capaç de generalitzar el suficient com per poder reconèixer les accions realitzades a diferent distància, angle de visió, etc.
- Segmentació d'accions: A la seva vida quotidiana les persones realitzem accions o gestos de forma continua concatenant una acció amb la següent. Per poder reconèixer les accions realitzades en una seqüència d'accions o gestos concatenats, els sistemes han de ser capaços d'extreure les accions individuals que formen el flux d'accions.

Per donar solució a cadascuna d'aquestes fases, els investigadors han proposat solucions que van des de l'aplicació de tècniques que deriven del reconeixement d'objectes [10, 11, 13], del reconeixement de la parla [42], del reconeixement facial [26] o del funcionament del cervell [8, 9].
Alguns autor han estès els descriptor/extractors de característiques del reconeixement d'objectes al domini espai-temporal amb la intenció d'aprofitar els bons resultats obtinguts en aquesta disciplina [5, 6, 7]. Altres han estès tècniques bioinspirades que també han donat resultats molt prometedors en el reconeixement d'objectes [14, 76, 77, 78]. Recentment, les tècniques d'aprenentatge profund (*deep learning*) han aconseguit molt bons resultats en diferents àmbits de la intel·ligència artificial. A l'àmbit de

la visió per computador, el *deep learning* ha estat principalment aplicat per les xarxes neuronals convolucionals (CNNs). Aquestes, han mostrat que poden superar als mètodes actuals en els reconeixement d'objectes i estan mostrant també molt bons resultats en el reconeixement d'accions humanes.

L'èxit aconseguit per les CNNs (*learned features*) i per les tècniques no-profundes (*engineered features*), ha fet que molts investigadors hagin intentat aprofitar les avantatges dels dos grups de tècniques aprofitant la complementarietat existent entre les característiques dels dos grups de tècniques. D'aquesta forma, les tècniques no-profundes encara continuen sent útils per funcionar de forma autònoma o bé per combinar amb tècniques profundes i millorar els resultats de l'estat de l'art.

Un darrer grup d'autors han plantejat tècniques que condensen seqüències de vídeo en una o varies *templates* que conserven informació discriminant. Alguns d'aquests autors han intentat entendre la geometria interna de les seqüències d'imatges considerant aquestes com a tensors multidimensionals [15, 16].

Les tècniques presentades a aquesta tesi, estan relacionades amb aquest conjunt de tècniques que intenten condensar una seqüència de vídeo en uns *templates* que retinguin informació important de cara a la discriminació entre accions/gestos humans, i aprofitar la complementarietat de les característiques extretes d'aquests *templates* amb altres característiques.

A la primera part d'aquesta tesi, hem proposat una aproximació basada en *template* per la representació d'accions/gestos a partir de tensors. Les nostres *templates* es calculen des de 3 projeccions diferents considerant una seqüència de vídeo com un tensor de tercer ordre. Calculem cada projecció des de les fibres del tensor de tercer ordre utilitzant funcions simples. Les fibres són subtensors que es creen fixant tots els índexs menys un. Les *templates* resultants, obtenen molt bon rendiment en reconeixement d'accions i gestos utilitzant descriptors de característiques en els keypoints estrets d'aquestes *templates*.

Hem fet un estudi exhaustiu per trobar quina funció ha de ser utilitzada per projectar el *template* des del tensor, i quin extractor/descriptor és el més adequat. Per fer això hem testejat cinc funcions diferents simples per projectar les fibres, anomenats, *Max*, *Mean*, *Standard Deviation*, *Kurtosi* i *Skewness*. Hem estudiat també el rendiment obtingut utilitzant quatre tècniques d'extracció/descripció de característiques de l'estat de l'art com PHOW, LIOP, HOG i SMFs en les nostres *templates*.

Per valorar el rendiment de la nostra tècnica, hem fet diversos experiments. A un dels experiments hem comparat la nostra aproximació amb *Motion History Images* (MHI) en tres simples *datasets* públics d'accions humanes. Utilitzant *datasets* més complexes, hem estudiat quina és la millor representació de les característiques extretes de les *templates* (*Bag Of Words* o *Fisher Vectors*) i la complementarietat entre les característiques extretes amb cada una de les cinc funcions (*Max*, *Mean*, *Standard Deviation*, *Kurtosi* i *Skewness)* i la complementarietat d'aquestes amb una exitosa tècnica com *Improved Dense Trajectories* (IDTs) [79].

Els experiments han demostrat que la desviació estàndard és la millor funció per projectar les fibres en les *templates*, i que PHOW obté el millor rendiment com a detector /descriptor en les *templates* obtingudes.

Els *datasets* més complexes han mostrat que la millor representació per a les característiques extretes de les *templates* és *Fv*, que existeix complementarietat entre les característiques extretes amb cada una de les funcions i que la fusió d'aquestes característiques amb IDTs, fa que aquest últim millori el seu rendiment.
Derivat dels treballs d'aquesta tesi, presentem una altre aproximació basada en *template* per reconeixement d'accions/gestos que s'obté d'una projecció derivada de la transformada de Radon i que permet la segmentació temporal d'accions en temps real. Primer plantegem una generalització de la transformada *R* que permet adaptar la transformada al problema a resoldre mitjançant la funció de projecció. Hem estudiat el seu rendiment per a les funcions *Max, Mean i Standard Deviation* en reconeixement d'accions sobre un *dataset* públic i comparat els resultats amb la transformada *R* utilitzant PHOW com a descriptor de característiques. També hem comparat la nostra tècnica amb altres tècniques de reconeixement d'accions de l'estat de l'art utilitzant el mateix *dataset*.
Totes les tècniques les hem avaluat comparant-les amb les tècniques de reconeixement d'accions/gestos de l'estat de l'art amb *datasets* públics actuals. Els resultats han mostrat que la funció *Max* obté el millor resultat quan s'aplica sobre la transformada de Radon i que la nostra tècnica supera a molts mètodes de l'estat de l'art en reconeixement d'accions/gestos.
A una segona fase, hem introduït una modificació a la etapa de classificació de la nostra tècnica per permetre segmentar accions temporalment. Per avaluar el seu rendiment, hem concatenat les 10 accions del Weizmann *dataset* i mesurat la capacitat de la tècnica per identificar cadascuna de les accions individuals. En aquest cas l'experiment té l'objectiu de valorar la capacitat de la tècnica per segmentar accions d'interès d'altres tipus de moviments de persones o objectes a l'escena, hem entrenat el nostre sistema considerant les accions del Weizmann *dataset* com a accions d'interès i les del Hollywood *dataset* com a accions de no interès.
Els experiments han demostrat que la nostra tècnica rendeix millor en la segmentació de les accions del Weizmann *dataset* que les tècniques de l'estat de l'art. També es demostra que tot i que el segon experiment no tenim constància de que ningú l'hagi fet abans, el rendiment és molt bo, obtenint-se un 100% de *Recogntion Rate* (*RR*).

1.1. Motivació

Els bons resultats en els darrers anys en el reconeixement d'objectes a imatges estàtiques, genera grans expectatives de poder utilitzar aquestes tècniques en el reconeixement d'accions/gestos. D'altre banda alguns sistemes bioinspirats han aconseguit molt bons resultats aplicant lògiques i funcions molt simples en el reconeixement d'objectes, però que poden tenir un alt cost computacional quan s'apliquen a seqüències d'accions/gestos. Això suposa una motivació per desenvolupar tècniques que puguin concentrar informació de tota una seqüència en pocs *templates* que permetin aplicar les tècniques més exitoses del reconeixement d'objectes. Idealment aquestes *templates* s'haurien de poder construir aplicant

funcions simples com les utilitzades a les tècniques bioinspirades, però amb baix cost computacional i que permetin l'execució en temps real.

1.2. Objectius

L'objectiu d'aquesta tesi és plantejar una nova tècnica de reconeixement d'accions/gestos que aporti millores a determinats aspectes no resolts amb les tècniques de l'estat de l'art, com poden ser:

- Permetre l'aplicació de tècniques exitoses de reconeixement d'objectes al reconeixement d'accions/gestos humans
- Reconeixement d'accions/gestos humans on pot existir auto-oclusió
- Segmentació temporal d'accions/gestos humans
- Alt rendiment i computació en temps real

Tot i que algunes tècniques actuals resolen alguns d'aquests problemes individualment de forma exitosa, molt poques vegada donen una solució global acceptable a tots aquests problemes. L'objectiu d'aquesta tesi és plantejar una tècnica de reconeixement d'accions/gestos humans que doni una solució global exitosa a aquests problemes.

1.3. Contribucions

En aquesta tesi, hem fet les següents contribucions:

- Nova tècnica per construir *templates* temporals a partir de les fibres de subtensors d'una seqüència d'accions aplicant funcions simples, que permet aplicar tècniques de reconeixement d'objectes de l'estat de l'art i que poden ser executat en temps real.
- Estudi del rendiment obtingut en els reconeixement d'accions/gestos humans per diferents funcions simples aplicades als subtensors.
- Estudi del rendiment obtingut en el reconeixement d'accions/gestos humans per diferents tècniques d'extracció/descripció i representació de característiques i de reconeixement d'objectes aplicades sobre les nostres *templates*.
- Estudi de la complementarietat de les característiques extretes de les noves *templates*.

- Mètode complet de reconeixement d'accions/gestos basat en les nostres *templates* que obté un alt rendiment, reduint el problema de l'oclusió, sent invariant a escalat i translació, sense necessitat de la detecció prèvia de keypoints i que permet ser computat en temps real.
- Proposta d'una nova transformada anomenada R_f que és una variant de la transformada R i que permet adaptar la transformada al problema a resoldre.
- Noves *templates* construïdes amb la transformada R_f, que permeten la classificació d'accions humanes i la segmentació temporal d'accions/gestos en seqüències d'accions concatenades en temps real.

2. Estat de l'art

A la primera part d'aquest capítol introduïm les tècniques generals de reconeixement de gestos humans. A la segona part s'introdueix el reconeixement d'accions humanes a seqüències d'imatges. A la tercera part, plantegem l'estat de l'art de les tècniques de segmentació d'accions humanes en seqüències de vídeo. L'objectiu no és fer un repàs de totes les tècniques utilitzades en aquesta disciplina, sinó més aviat plantejar les tècniques que han estat més exitoses en els darrers anys i que justifiquen el nostre treball de recerca a aquesta tesi.

2.1. Reconeixement de gestos

El reconeixement gestual de mans proporciona una atractiva alternativa a incòmodes interfases per a la interacció home-màquina. Aquest fet ha motivat una recerca molt activa basada en la visió per computador, i centrada en l'anàlisi i interpretació dels gestos de mans. Molts sistemes de visió per computador i reconeixement de formes, incloent; extracció de característiques, detecció d'objectes, *clustering*, i classificació, s'han utilitzat de forma exitosa en el reconeixement gestual [80, 81]. Treballs preliminars d'interpretació gestual basats en visió, van ser enfocats en el reconeixement de postures o gestos de mans estàtics. Però, els gestos de mans són accions dinàmiques i el moviment de mans porta més informació que les postures. Mentre que el reconeixement gestual (postures) pot ser realitzat per tècniques de *matching template* i reconeixement de formes, el reconeixement de gestos dinàmics involucra l'ús de de tècniques con *Dynamic Time Warping* (DTW) [82] o *Hidden Markov Models* (HMM) [83, 84].

Ja que hi ha bastant similitud entre les tècniques utilitzades al reconeixement gestual i de la parla [82], les tècniques com HMM o DTW, generalment utilitzades en reconeixement de la parla, són també utilitzades en el reconeixement gestual. A les aplicacions de reconeixement de la parla, el reconeixement de paraules, independent de la seva durada i variacions en la pronunciació, ha quedat provat con un problema molt complicat. Els HMMs han mostrat que resolen aquest problema de forma exitosa. Un HMM s'associa amb cada unitat d'idioma diferent, mentre que en reconeixement gestual cada gest pot ser associat amb un diferent HMM.

DTW i HMM han estat aplicats en una gran quantitat de treballs relacionat amb el reconeixement gestual, però cap d'aquest treballs ha fet una comparativa exhaustiva d'aquestes dues tècniques. A [85], vam fer un estudi amb profunditat del rendiment d'aquestes dues tècniques.

Tot i que el reconeixement de gestos continua sent una disciplina de molt d'interès entre els investigadors de la visió per computador, en els últims anys el reconeixement d'accions humanes ha guanyat importància com una forma més general del reconeixement de gestos. A l'àmbit del reconeixement d'accions humanes ha hagut un gran avanç, considerant tècniques molt més avançades que les aplicades al reconeixement de gestos, és per això que considerem que és més important estendre aquesta disciplina més general i que comentem al següent apartat.

2.2. Reconeixement d'accions humanes

Actualment el reconeixement d'accions humanes és una important àrea de recerca de la visió per computador on l'objectiu és detectar i reconèixer automàticament accions humanes d'una seqüència de vídeo.

Una acció humana a una seqüència es pot considerar com un objecte 4-dimensional, que es pot descompondre en les seves parts espacials i temporals.

D'altre banda, el reconeixement d'objectes en imatges tracta de classificar un objecte en una categoria prèviament definida. Aquesta disciplina ha aconseguit molt d'èxit en el últims anys. Una de les raons d'aquest èxit, ha estat la robustesa dels extractor/descriptors de característiques utilitzats, com *Scale Invariant Feature Transform* (SIFT) *algorithm* [17], *Histogram Of Oriented Gradients* (HOG) [18], *Local Intensity Order Pattern* (LIOP) [19] o variants d'aquestes tècniques com *Pyramid Histogram Of visual Words* (PHOW) [20]. Aquests èxits han motivat a molts autors a utilitzar aquestes tècniques per el reconeixement d'accions, estenent aquestes tècniques a vídeo seqüencies. A [6] va utilitzar 3D SIFT per reconeixement d'accions utilitzant característiques espai-temporals. També HOG va ser estès a vídeo seqüències [5]. Els autors van calcular histogrames 3D en les orientacions dels gradients espai-temporals, i van aplicar-ho al reconeixement d'accions. SIFT i HOG van ser combinats amb el paradigma *Bag Of Words* (BoW) [86] i aplicats de forma exitosa al reconeixement d'accions a [87]. A [113], van estendre *Local Binary Patterns* (LBP) a *Volume Local Binary Patterns* (VLBP), combinant moviment i aparença però en aquest cas els van aplicar per reconèixer textures dinàmiques. Per fer VLBP computacionalment simple i fàcil d'estendre, només van considerar les co-ocurrències en tres plans separats. Inspirat en el descriptor *Shape Context*, [7] van introduir el *Motion Context* per capturar les estructures de moviments relatius. Tot i que aconsegueixen bons resultats, aquestes tècniques no consideren informació espai-temporal de forma global. El seu gran inconvenient és la dependència dels punt d'interès triats

Uns altres tipus de treballs s'han basat en el funcionament del cervell i aplicat aquestes idees al reconeixement d'objectes. A [8, 9] van proposar una tècnica bioinspirada anomenada *Standard Model Features* (SMFs), que intenta resumir un nucli de fets acceptats sobre el *stream* ventral al còrtex visual. A

la seva forma més simple, el model consisteix en quatre capes d'unitats computacionals, on les unitats simples *S* alternen amb les unitats complexes *C*. Les unitats *S* combinen les seves entrades amb una funció sintonitzada amb forma de campana per incrementar la seva selectivitat. Les unitats *C* filtren les seves entrades a traves d'una operació de màxim, incrementant la seva invariància. Motivats per la similitud a l'organització de les vies de la forma i el moviment en el còrtex visual, a [14] van intentar aplicar al reconeixement d'accions, mecanismes computacionals que han estat provats útils al reconeixement d'objectes. Els autors van utilitzar una arquitectura similar a la utilitzada al reconeixement d'objectes, però van afegir unes unitats específiques *S3* i *C3* pel processat del moviment i incloure invariància temporal. També van comparar tres tipus diferents d'unitats *S1*. Tot i que els *recognition rates* obtinguts a aquesta aproximació són molt bons, té el desavantatge del seu gran temps de processat (aproximadament 2 minuts per seqüència) que impedeix el seu ús en temps real. També van concloure que les etapes S3 i C3 incloses mostren una petita millora en dos dels *datasets* utilitzats i un empitjorament en un tercer *dataset*. La figura 1 mostra el diagrama de blocs de [14].

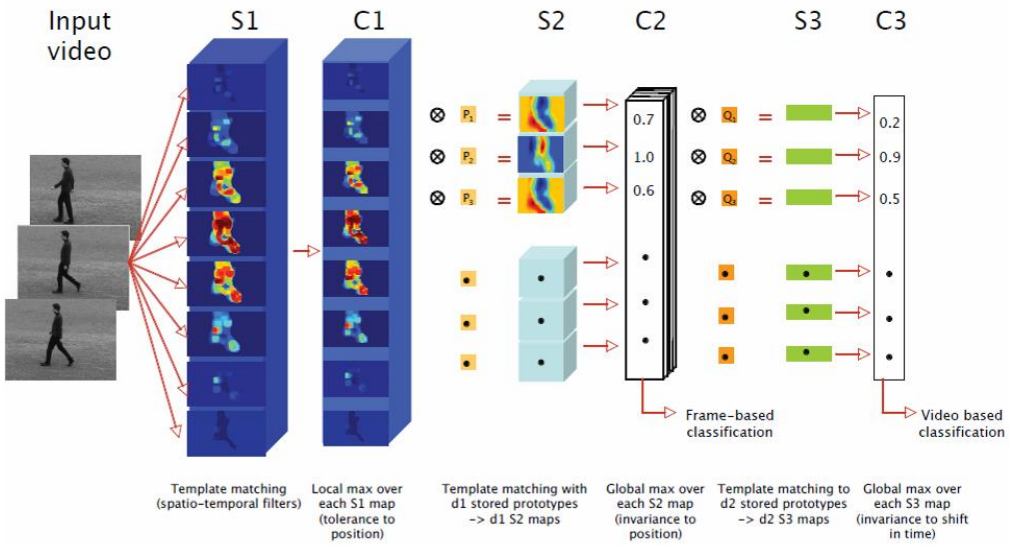


Figura 1. Diagrama de blocs dem **[14]**

D'altra banda, alguns autors suggereixen algunes extensions d'aquest model [88], tal que una altre part del sistema visual, com el *stream* ventral del còrtex visual involucrat en el anàlisi de la forma, pot ser important pel reconeixement de moviment [89].

Recentment, molt autors han començat a considerar les seqüències de vídeo com a estructures multidimensionals amb una representació subjacent de baixa dimensionalitat.

L'objectiu de les tècniques basades en *Manifold Learning* és aprendre automàticament aquesta representació. Si les dades originals s'estenen en un espai *Rd,* llavors el problema del *Manifold Learning* és trobar les coordenades de cada punt en un espai *d* de dimensió menor.

El problema de la reducció no lineal de la dimensionalitat ha estat un àrea de gran interès durant moltes dècades. *Isometric feature mapping* (Isomap) [21], *Local Linear Embedding* (*LLE*) [22] i *Stochastic Neighbor Embedding* (*SNE*) [23] són alguns dels algorismes que han estat proposats amb aquest objectiu. A [24], van comparar els algorismes *Isomap, LLE* and *t-SNE* (*t-Distributed Stochastic Neighbor Embedding*) i van concloure que *Isomap* és superior preservant més informació global de la relació entre els punts. També va proposar una alternativa al *SNE* anomenada *t-SNE*, capaç de superar l'estat de l'art en visualització i reducció de la dimensionalitat. Moltes tècniques de *manifold learning* basades en *Isomap* i *LLE* necessiten una gran quantitat de dades d'entrenament i un dens mostreig en el *manifold*. A les aplicacions reals moltes vegades no és possible disposar de totes aquestes dades.

Basat també en la idea de *manifolds,* [25] va estendre un descriptor de forma com la transformada *R* [27] per descriure accions. La transformada *R* és computacionalment eficient i robusta per moltes transformacions comunes d'imatges. Aquesta transformada converteix una silueta d'una imatge en un senyal compacte *1D* a través de la transformada de Radon. La transformada *R* estén la transformada de Radon calculant la suma dels quadrats dels valors de la transformada de Radon de totes les línies d'un mateix angle. A [25] van presentar un *framework* per aprendre una representació invariant al punt de vista d'una sèrie d'accions primitives (caminar, cop de puny, cop de peu, seure) obtingudes d'una única càmera. Aquest treball està relacionat amb [28], però en comptes d'aprendre un conjunt arbitrari de funcions base lineals, van modelar el canvi en aparença d'una acció deguda al punt de vista con a un *manifold* de baixa dimensionalitat. El principal inconvenient d'aquest treball és la necessitat d'extreure la silueta de la persona prèviament a l'aplicació de la transformada *R*.

Una altre grup d'autors, intenten explorar les característiques d'espai i realitzar la classificació basada en la geometria intrínseca del *manifold*. Aquests treballs consideren la seqüència de vídeo com un tensor de tercer ordre. Alguns d'aquests treballs estan basats en l'espai tangent [26, 30].

A [15] es va introduir el concepte de *Grassmann Manifold* en el context del reconeixement d'accions. Els autors, encasten l'espai tangent aproximat en un *Grassmann Manifol* on cada element en aquest *manifold* representa un subespai. L'aspecte clau d'aquest encastament és l'ús de la distància geodèsica, que està molt ben definida en els *Grassmann manifolds,* i per tant és considerada la geometria subjacent. Per fer això, ells representen una seqüència de vídeo com un tensor de tercer ordre i el despleguen en tres matrius per intentar caracteritzar l'estructura subjacent. La figura 2 mostra un exemple del desplegament del tensor.

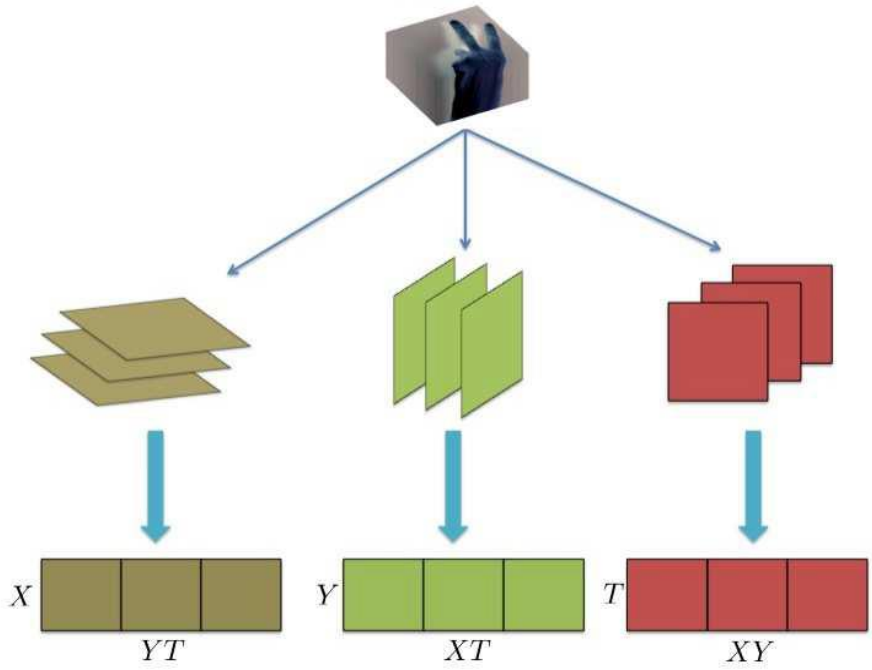


Figura 2. Un exemple de desplegat d'una matriu per un tensor de tercer ordre [15].

A [16] es presenta un *framework* estadístic per extreure característiques de similitud entre dos vídeos per reconeixement de gestos/accions. Els autors van estendre l'anàlisi clàssic de correlació canònica a un *array* de dades multidimensional per analitzar la similitud de volums de dades de vídeo espai-temporals. Aquesta extensió la van fer considerant el vídeo un tensor multidimensional i desenvolupant un *Tensor Canonical Correlation Algorithm* (TCCA). A [29], van utilitzar característiques SIFT amb CCA (*Canonical Correlation Analysis*) per reconeixement de gestos. A [31], van crear un *Pseudo Kernel Riemannia* i el van emprar per encastar un *Rienmanian Manifolds* en un *Reproducing Kernel Hilbert Space* (RKHS). Ja que CCA no considera la correlació no lineal entre característiques i amb l'objectiu de reduir la dimensionalitat de les característiques visuals extretes, [112] va proposar *Hessian Multiset Canonical Correlations* (HesMCC) per reducció de dimensionalitat en múltiples vistes. HesMCC treu profit de la geometria local intrínseca de les dades del *manifold* en contrast amb el Laplacià. A [111], van desenvolupar un *General Tensor Discriminant Analysis* (GTDA) com un pas de pre-processat per Anàlisi Lineal Discriminat (LDA) per reconeixement postural. Per representar les postures, van desenvolupar tres representacions d'imatges basades en funcions *Gabor*, anomenades, *GaborD*, *GaborS* i *GaborSD*. *GaborD* respon en funció de la direcció de la informació, *GaborS* respon en funció de l'escala de la informació i *GaborSD* respon a les dues informacions anteriors.

Tot i que aquestes tècniques han mostrat resultats molt exitosos, el seu gran inconvenient és la alta sensibilitat a canvis d'escala o a translacions.

La importància de modelar la dinàmica temporal de les activitats humanes va ser demostrada per [32]. Els autors van presentar un mètode de reconeixement que modela l'activitat com una sèrie temporal de histogrames no euclidians de característiques orientades de *Optical Flow* (OF).

Les trajectòries de vídeo també han estat utilitzades com una representació per l'anàlisi de vídeo i reconeixement d'accions. Basat en grafs encastats, [33] va modelar vídeo clips de gestos com trajectòries en el espai d'aparença escalada.

En els últims anys, l'intent d'aconseguir millor rendiment en *datasets* més complexes i realistes ha fet que, aprofitant també la millora del poder computacional actual, s'hagi produït un gran avanç de les tècniques de reconeixement d'accions humanes. Alguns d'aquests avanços s'han basat en la combinació de múltiples característiques extretes amb diferents descriptors. A [34], van proposar mostrejar punts de característiques a una *grid* densa de cada *frame* i fer un *tracking* d'ells utilitzant OF per millorar unes altres tècniques, con el KLT *tracker*. Van mostrejar punts de característiques a diferents escales i després aquests punts de característiques van ser seguits a cada escala separadament. Els punts de cada *frame* són concatenats per formar trajectòries. Diferents característiques com *Trajectories*, HOG, HOF i MBH [35] són combinades dintre de volums d'espai-temps alineades amb aquestes característiques utilitzant *spatio-temporal pyramids* (DTs). HOG captura l'aparença estàtica, HOF representa moviment, i MBH representa informació de moviment als contorns. MHB separa les components verticals i horitzontals de l'OF, es calculen les derivades espacials de les dues components i la informació d'orientació es quantifica en dos histogrames. La magnitud de l'OF s'utilitza per a ponderar.

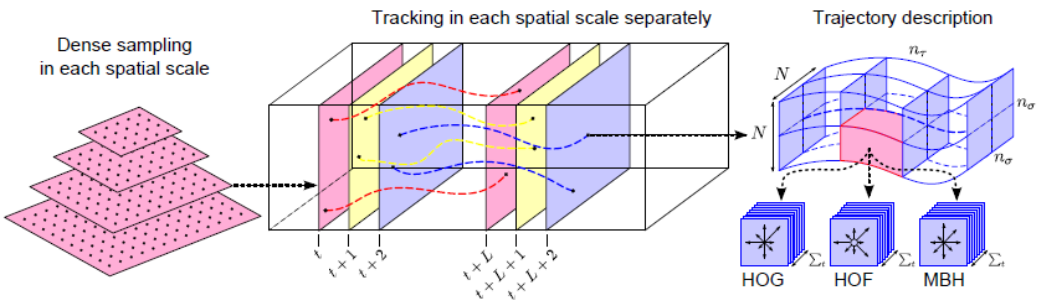


Figura 3. Exemple de [34] per extreure i caracteritzar les trajectòries denses.

A [79], presenten una millora de [34] anomenada *improved dense trajectories* (IDTs) per reconeixement d'accions humanes. Per estimar el moviment de càmera, fan un *match* dels punts de característiques utilitzant el descriptor SURF i OF dens. També milloren l'estimació utilitzant un detector de forma humana. IDTs millora el rendiment respecte a DTs considerant la correcció del moviment de la càmera.

Peng et al.[36] va avaluar la fusió de HOG, HOF i MBH a tres diferents nivells anomenats; nivell de descriptor, nivell de representació o nivell de *score*. El nivell de descriptor és realitzat al nivell de cuboids, concatenant múltiples descriptors des de cada cuboid a un únic descriptor. Al nivell de representació, la fusió es realitzada en el nivell de vídeo entrant els descriptors en un BoF separadament, i fusionant la representació global com un únic descriptor. Al nivell de *score*, la fusió es realitzada en el nivell de vídeo també, però cada descriptor entrena un classificador i al final la classificació és obtinguda fusionant els *scores* de múltiples classificadors. També van avaluar diferents codificadors de característiques com *Vector Quantization Encoding* (*VQ*) [37], *Localized Soft Assignment Encoding* (*SA-k*) [38], *Locality-constrained Linear Encoding* (*LLC*) [39], *Fisher Vector* (*Fv*), *Vector of Locally Aggregated Vector* (*VLAD*) [40] i *Super and Vector Coding* (*SVC*) [41] i van obtenir els millors resultats utilitzant la codificació *Fv* amb fusió a nivell de representació.

Peng et al.[43] van extreure les característiques de les trajectòries millorades de [79] i les van codificar fent servir *Fv* i *Vector of Locally Aggregated Descriptors* (VLAD) separadament. Finalment van concatenar *Fv* i VLAD en un *Hybrid Super Vector*. Wu et al. A [44] van avaluar diferent millores per codificació de vídeo basades en VLAD i van suggerir fer servir VLAD o *Fv* en comptes de BoW. Peng et al. havia també proposat a [36] una representació de característiques simples anomenada representació híbrida, explorant la complementarietat de diferents models de *Bag* of *Visuals Words frameworks* i descriptors locals.

Recentment, Hyun-Joo et al.[45] va proposar un nou model anomenat *Bag-of-Sequencelets* (BoS). Un model BoS representa un vídeo com una seqüència d'accions primitives (PA), considerant una acció complexa com una composició ordenada de sub-accions. Per permetre aprenentatge PA, van representar una vídeo seqüència con una seqüència de característiques IDTs i representació codificada com *Fv*.

Un grup de tècniques que han obtingut molt bon resultats són les *Convolutional Neural Networks* (CNNs). Les CNNs són tècniques bioinspirades que han estat esteses al reconeixement d'accions humanes [76, 77 ,78]. Ng et al.[76] van proposar un *Pooling* de característiques i xarxes neurals recurrents per combinar informació de la imatge per gestionar vídeos llargs. Van explorar també la necessitat d'informació de moviment i van confirmar que pel *dataset* UCF101 és necessari fer servir el OF per obtenir resultats acceptables. A la seva aproximació van combinar informació espacial i temporal. A més, van mostrar que *Long Short-Term Memory* (LSTM) sobre un *Pooling* temporal de característiques no millora (o millora només marginalment) els resultats depenent del *dataset* utilitzat.

Simonyan et al. [77] també va combinar informació espacial i temporal utilitzant una arquitectura basada en dos fluxos separats (espacial i temporal) combinats amb fusió tardana. Van establir que fent servir OF com a entrada, el seu model no requereix un significant *hand-crafting*, però diuen també que els ingredients essencials de la representació superficial (No *deep*) com el *Pooling* de característiques locals sobre tubs espai-temporals, són perduts a la seva arquitectura.

Zhichen et al. [46] van proposar un algorisme d'aprenentatge semàntic per reconeixement d'accions a partir d'imatges fixes. Van definir la "part semàntica" com qualsevol regió que proveeix una gran contribució al correcte reconeixement. Van utilitzar *Pooling* a CNNs per trobar objectes interactius i informació de postures i després van combinar ambdós per formar una representació més discriminativa.

Darrerament, alguns autor han observat que un dels problemes que fan baixar el rendiment a les CNNs és que la majoria de mètodes només consideren els vídeos per fragments, sense considerar una representació global d'aquest, de forma que molts errors de classificació vénen donats per accions humanes que tenen parts comunes i que tractades per fragments no poden ser diferenciades. Ionut et al. [117] van proposar un mètode de codificació de característiques específicament dissenyat per característiques profundes locals anomenat *Spatio-Temporal Vector of Locally Max Pooled Features (ST-VLMPF)*, que captura una representació global de la vídeo seqüència. Rohit et al. [116] van presentar una nova representació per classificació d'accions que agrega característiques convolucionals locals a través de tota l'extensió espai-temporal del vídeo. El nucli d'aquesta representació és una nova capa anomenada *ActionVLAD* que és una extensió espai-temporal de la capa d'agregació *NetVLAD*. Varol et al. [114] van fer servir CNNs amb convolucions temporals *long-term* per representar les accions en la seva total extensió. Ells incrementen l'extensió temporal de la representació a costa de reduir la resolució espacial per conservar la complexitat de la xarxa tractable. També demostren que la qualitat del OF utilitzat afecta de forma significativa al rendiment final obtingut. Yunbo et al. [119] afirma que tractar diferents *streams* independentment i després fusionar-los mitjanant els valors no és suficient, ja que a la majoria dels casos d'error de classificació un dels *streams* falla mentre que la resta son correctes. Per solucionar aquesta problemàtica, van proposar una xarxa espai-temporal piramidal per fusionar les característiques temporals i espacials. A [115], els autors van fer una extensió de [78] considerant informació estàtica, *Dynamic Images*, OF i *Dynamic Optical Flow*, i van combinar els quatre *streams* mitjanant els *scores* obtinguts de cada *stream*. Feichtenhofer et al. [118] va presentar una arquitectura espai-temporal creada amb interacció multiplicativa de característiques d'aparença i moviment.

Amb l'objectiu de reduir la informació a processar en seqüències d'imatges, Bobick&Davies van introduir *templates* temporal a [1]. Aquests consisteixen en imatges 2D calculades des de seqüències d'imatges, que retenen informació temporal important. En aquest treball van presentar dos *templates* temporals anomenades *Motion Energy Images* (*MEI*) i *Motion History Images* (MHI). MEI és una imatge binaria que codifica regions de la imatge on ha hagut moviment i MHI és una imatge de nivells de gris que codifica com de recent és el moviment ocorregut en un píxel. Un dels inconvenients d'aquesta aproximació és el problema de l'auto-oclusió. Les accions que contenen més d'una direcció són anomenades accions complexes. Una acció complexa típica és l'acció de seure i aixecar-se, ja que està formada pel moviment de seure i el d'aixecar-se que tenen diferents direccions. L'auto-oclusió succeeix perquè els darrers moviments sobreescriuen els moviment previs a la mateixa regió. A [2] van utilitzar una estructura compartida jeràrquica per reconèixer accions complexes. A [4] van utilitzar una gramàtica

lliure de context per representar accions complexes. A [12] els autors van modelar seqüències temporals de les postures del cos estimada a cada *frame*. A [47] es va proposar una millora del problema d'auto-oclusió de MHI i MEI calculant el *OF* i separant aquest en quatre direccions diferents. A les seves conclusions reconeixen que els resultats no són massa bons i que han de continuar treballant en el mètode, a més no mostren resultats d'aquesta aproximació en cap *dataset* dels normalment utilitzats. També amb l'objectiu de solucionar aquest inconvenient, a [109] van introduir *Motion History Histograms* (MHH), el qual manté informació del número de vegades que un moviment és detectat en cada píxel, i a més categoritzat en el llarg de cada moviment. [110] va proposar un MHI/MEI ponderat que utilitza funcions *fuzzy* per emfatitzar la informació de moviment en diferents regions temporals en comptes de l'últim *frame* com el MHI tradicional.

A [78], els autors van presentar una nova representació de vídeo anomenada *Dynamic Images* que condensa una seqüència de vídeo a una única imatge fent servir *rank pooling* i CNNs. Van proposar també una capa de *Pooling* temporal estenent les *Dynamic Images* a un mapa de característiques CNNs.

2.3. Segmentació temporal d'accions

Tot i que la majoria de la literatura actual sobre reconeixement d'accions/gestos es basa principalment en el reconeixement d'accions prèviament segmentades temporalment, a les aplicacions del món real les accions o gestos es succeeixen consecutivament en el temps sense cap identificatiu d'inici o final.

A [48] es presenta una tècnica que parametritza la sortida de probabilitats subjacent dels estats de un HMM. Els autors estenen els HMM estàndard per incloure una variació de paramètrica global en la sortida dels estats del HMM (PHMM) i formula un mètode d'*Expectation-Maximization* (EM) per entrenar el PHMM.

A [49] intenten segmentar els gestos humans en moviments atòmic detectant punts d'inconsistència en l'observació global com variacions abruptes en la direcció dels moviments, i trobant mínims locals en la velocitat i màxim locals en els canvis de direcció. Després clusteritzen els segments resultant d'aquest procés utilitzant HMM per evitar la sobre-segmentació. D' aquesta forma el resultat del *clustering* dóna una representació de l' observació contínua original, en la que cada segment és reemplaçat pel número de clúster al que pertany. Per provar el seu sistema els autors van gravar 8 minuts de gestos de direcció musical i afirmen que la seva aproximació depèn de la freqüència dels gestos i que és adequada per moviments estructurats humans que es repeteixen moltes vegades.

A [50] van representar les accions humanes com a seqüències curtes de postures atòmiques del cos que guarden com un conjunt de siluetes des de diferents punts de vista (multicàmera). Van extreure les accions de les postures atòmiques constituents d'un conjunt de vídeo seqüències i les van utilitzar per

construir un HMM. La seqüència de postures es *parsejada* per trobar les accions, i aquestes són descrites per una gramàtica a la qual les parelles de postures atòmiques actuen com a símbols terminals. Els autors comenten algunes limitacions del seu sistema, com per exemple, que descriu les accions utilitzant postures estàtiques.

A [42] també es va utilitzar un HMM, en aquest cas inspirant-se en el reconeixement de la parla, van dividir el problema d'inferència en dos nivells. El nivell més baix proposa candidats de detecció basat en característiques de baix nivell fent servir un HMM. Les sortides d'aquest detectors serveix com entrada a un mecanisme de gramàtica lliure de context estocàstic de *Parsing*. El principal inconvenient d'aquesta tècnica és el cost de realitzar inferència gramatical en un *Framework* de *Parsing* estocàstic. A [51] consideren un esdeveniment com procés temporal estocàstic i agafen les característiques locals en múltiples escales temporals com mostres del procés estocàstic. Més tard, els utilitzen per construir una distribució empírica associada amb l'esdeveniment. Per aïllar i clusteritzar els esdeveniments, mesuren la distància entre les distribucions per obtenir una distància estadística entre vídeo seqüències. Els autors, plantegen que aquesta mesura pot no ser òptima per una acció específica, però permet un anàlisi general basat en esdeveniments d'informació de vídeo que contenen tipus d'esdeveniments desconeguts. La figura 4 mostra un exemple de la distribució empírica.

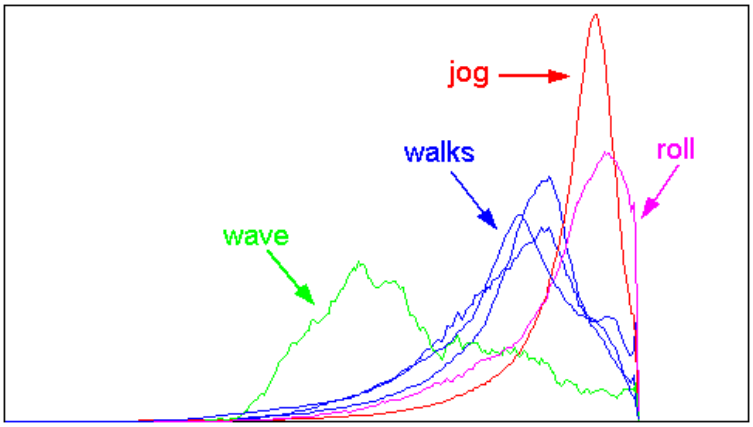


Figura 4. Exemple de distribució empírica de 6 diferents clips [51].

A [52] mostren que la minimització de l'entropia de la component de les distribucions i l'estat intern d'un HMM pot ser construït per organitzar l'activitat observada en estats alts ocults interpretables, que capturen la regularitat dinàmica del conjunt d'entrenament. Afirmen que els estats ocults descoberts no garanteixen la coincidència amb els esdeveniments d'interès, però que per la seva experiència aquests sempre han estat interpretables i útils. A [53], van intentar segmentar activitats d'alt nivell en les seves sub-accions utilitzant HMM [54] modificats per gestionar dades absents en el vector d'observació. Els autors conclouen que tot i que els HMMs no són ideals per modelar la relació jeràrquica entre activitats i

accions, el mètode és encara capaç de segmentar els límits de les accions amb precisió. A [55], proposen una aproximació basada en l'extensió del *Conditional Randon Fields* (CRFs) [56] i models de Markov de màxima entropia [57] per reconeixement de moviment humà. Realitzen la inferència utilitzant programació dinàmica, mentre que l'entrenament està basat en un problema convex que garanteix un òptim global. A [58], proposen un *Framework* unificat que codifica relacions espai-temporals entre parts en moviment de l'aparença de postures individuals. Fan servir un algorisme d'aprenentatge de moviment basat en parts no supervisat.

Altres autors han utilitzat classificadors SVM, a [59] plantegen dos mètodes per segmentar moviments humans periòdics en cicles temporals. Apliquen el descriptor de característiques basat en formes *Pyramid Correlogram of Oriented Gradients* (PCOG) sobre MHI i MEI per capturar informació de moviment i forma. Finalment, utilitzen un classificador SVM multiclasse amb un *kernel* RBF. A [60], es proposa un mètode *One-Shot* basat en descriptors *3D Histograms of Scene Flow* (3DHOFs) i *Global Histograms of Oriented Gradient* (GHOGs) sobre imatges RGBD. Per la classificació utilitzen els *scores* de les SVMs entrenades per cadascuna de les classes. Afirmen que la tècnica pot funcionar en temps reals però necessita el seguiment de les parts del cos. La figura 5 mostra un diagrama de funcionament d'aquest mètode.

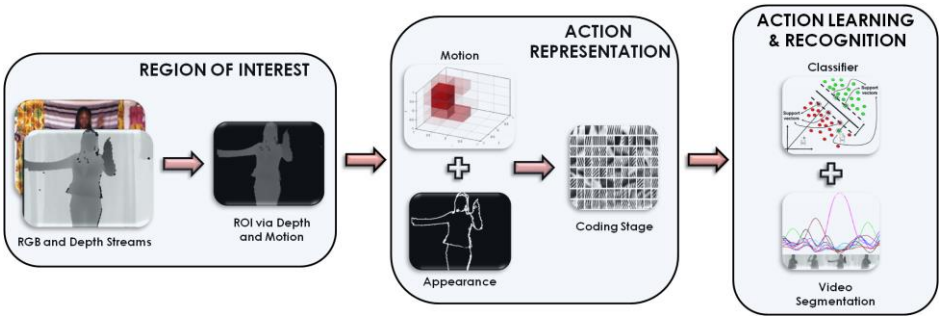


Figura 5. Diagrama de blocs de la tècnica [60].

A [61], proposen un mètode ajuntant segmentació de vídeo i reconeixement d'accions. Els autors, fan una extensió de BOW espacial amb un classificador SVM multiclasse, mentre que la inferència sobre els segments la fan amb programació dinàmica maximitzant els *scores* de les SVMs de la classe guanyadora, i suprimint la resta de classes.

Alguns autors s'han basat en característiques volumètriques per detectar esdeveniments. A [62], van proposar un mètode per detecció d'esdeveniments en entorns real dinàmics de multituds basat en les formes dels volums espai-temporals. Els esdeveniments són detectats utilitzant un descriptor de forma volumètric presentat per ells en combinació amb el descriptor de flux de Shechtman i Irani [63]. Més tard, trenquen les *templates* de l'acció en parts. La principal limitació d'aquest treball és la seva manca de

capacitat de generalització, ja que el model s'extreu d'un únic exemplar de l'esdeveniment. A [64], es proposa un *Framework* de característiques volumètriques per analitzar el vídeo que és una extensió del treball de Viola i Jones per detecció d'objectes en imatges estàtiques al domini espai-temporal. Els autors mostren un exemple de limitació en el reconeixement de dos accions similars degut a que el sistema es basa només en característiques de moviment, i proposen com a solució afegir un model d'aparença per incrementar la precisió. A [65], utilitzen un nou descriptor de moviment anomenat *Motion History Volume* (MHV) [66], el qual sintetitza el contingut d'accions d'una seqüència multivista curta, sense coneixement de les parts del cos. Més tard clusteritzen els MHVs resultants en una jerarquia de classes que permet reconèixer múltiples ocurrències d'accions repetides. Com a exemple del seu resultat utilitzen accions simples realitzades per dos membres del seu laboratori.

A [67], es proposa una funció de correspondència per mesurar la interdependència mútua i d'aquesta forma detectar esdeveniments i característiques importants simultàniament. Els autors extreuen els objectes d'interès de cada *frame* del vídeo sense fer *tracking* posteriorment dels objectes. Després, calculen els histogrames espacials dels objectes detectats i apliquen quantificació vectorial classificant-los en un diccionari de *K* característiques prototip. A [61], presenten una tècnica per capturar canvis abruptes utilitzant la curvatura espai-temporal de trajectòries 2D. Afirmen que aquesta representació és compacta, invariant a vista, i és capaç d'explicar una acció en termes d'unitats d'acció anomenades *dynamic instants and intervals*, on un *dynamic instants* és una entitat instantània que succeeix en un sol *frame* i representa un canvi important a les característiques del moviment. Un *interval* en canvi, representa un període de temps entre dos *dynamic instants* durant el qual les característiques del moviment no canvien.

3 Templates temporal a partir de subtensors

3.1 Introducció

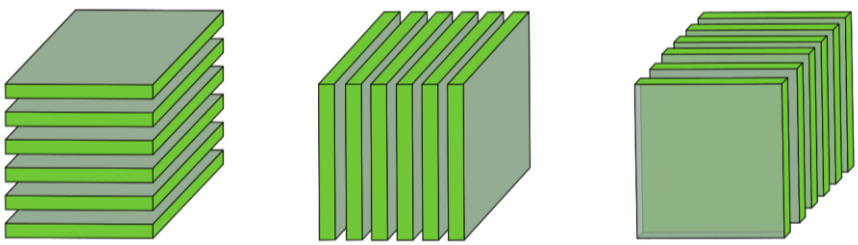
A aquest capítol presentem una de les aportacions d'aquesta tesi. Tracta d'una tècnica que redueix una seqüència d'imatges a una o més imatges mitjançant tres projeccions de la seqüència d'imatges original, considerant aquesta com a un tensor de tercer ordre.

3.2 Definició de tensor

Un tensor de tercer ordre pot ser considerat com un conjunt de matrius anomenades *Slices*.

De forma semblant a [62, 63], nosaltres utilitzarem a_{ij} per representar l'element *(i, j)* d'una matriu *A* i x_{ijk} per representar l'element *(i,j,k)* del tensor de tercer ordre *X*. $A^{(n)}$ representa la *n*th matriu d'un tensor de tercer ordre.

Els *Slices* poden ser horitzontals, verticals i frontals. Els *Slices* són definits fixant tots els índexs menys dos, i són denotats per $X_{i::}, X_{:j:}$ i $X_{::k}$ respectivament, on els dos punts indiquen tots els elements d'un mode. La figura 6 mostra els *Slices* d'un tensor de tercer ordre.



(a) Horitzontal *Slices*: $X_{i::}$ (b) Vertical *Slices* $X_{:j:}$ (c) Frontal *Slices* $X_{::k}$

Figura 6. *Slices* per a un tensor de tercer ordre.

Un tensor pot ser també considerat com un conjunt de vectors anomenats fibres. Una fibra es defineix fixant cada índex excepte un. El tensor de tercer ordre té fibres columnes, fibres files i fibres profunditat

denotades per $x_{:jk}$, $x_{i:k}$ i $x_{ij:}$, respectivament, on els dos punts indiquen tots els elements d'un mode. Per tant, el *mode 1* és el subespai representat per les fibres verticals, el *mode 2* és el subespai representat per les fibres horitzontals i el *mode 3* és el representat per les fibres de profunditat. La figura 7 mostra les fibres d'un tensor de tercer ordre.

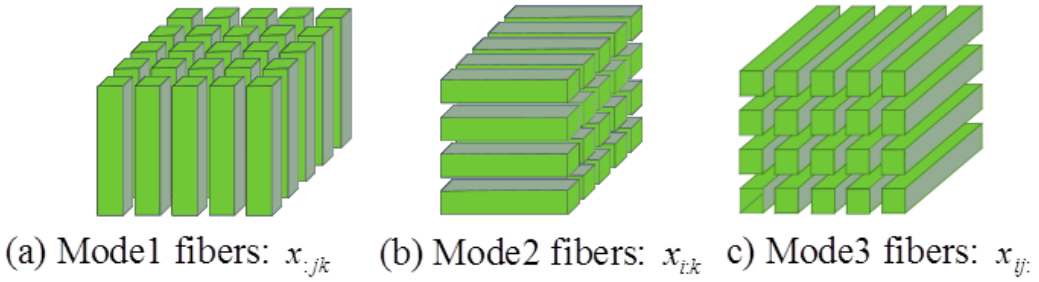


Figura 7. Fibres per a un tensor de tercer ordre.

3.2.1 Definició de tracte

Definim un tracte com un conjunt de fibres veïnes. Les equacions 4, 5 i 6 mostren els tractes dels *modes 3*, *mode 2* i *mode 1* denotades per T_{xy}, T_{xz} i T_{yz} respectivament.

$$T_{xy}^{\beta}(i,j) = \bigcup_{r=i-\frac{\beta}{2}}^{i+\frac{\beta}{2}} \bigcup_{s=j-\frac{\beta}{2}}^{j+\frac{\beta}{2}} (x_{rs:}) \qquad (4)$$

$$T_{xz}^{\beta}(i,j) = \bigcup_{r=i-\frac{\beta}{2}}^{i+\frac{\beta}{2}} \bigcup_{s=j-\frac{\beta}{2}}^{j+\frac{\beta}{2}} (x_{r:s}) \qquad (5)$$

$$T_{yz}^{\beta}(i,j) = \bigcup_{r=i-\frac{\beta}{2}}^{i+\frac{\beta}{2}} \bigcup_{s=j-\frac{\beta}{2}}^{j+\frac{\beta}{2}} (x_{:rs}) \qquad (6)$$

Sent β l'amplada del veïnat de les fibres que formen el tracte. Les equacions 7, 8 i 9 són utilitzades per calcular les projeccions, on les funcions $f_1\ R^{\beta\times\beta\times K} \to R$, $f_2\ R^{\beta\times\beta\times J} \to R$, i $f_3\ R^{\beta\times\beta\times I} \to R$ són aplicades a cada tracte T_{xy}, T_{xz} i T_{yz} .

$$I^{\beta}_{xy}(i,j) = f_1\left(T^{\beta}_{xy}(i,j)\right) \qquad (7)$$

$$I^{\beta}_{xz}(i,j) = f_2\left(T^{\beta}_{xz}(i,j)\right) \qquad (8)$$

$$I^{\beta}_{yz}(i,j) = f_3\left(T^{\beta}_{yz}(i,j)\right) \qquad (9)$$

$I_{xy(i,\,j)}$, $I_{xz(i,\,j)}$, i $I_{yz(i,\,j)}$ són píxels amb coordenades *i, j* de la projecció *XY*, projecció *XZ* i projecció *YZ* respectivament.

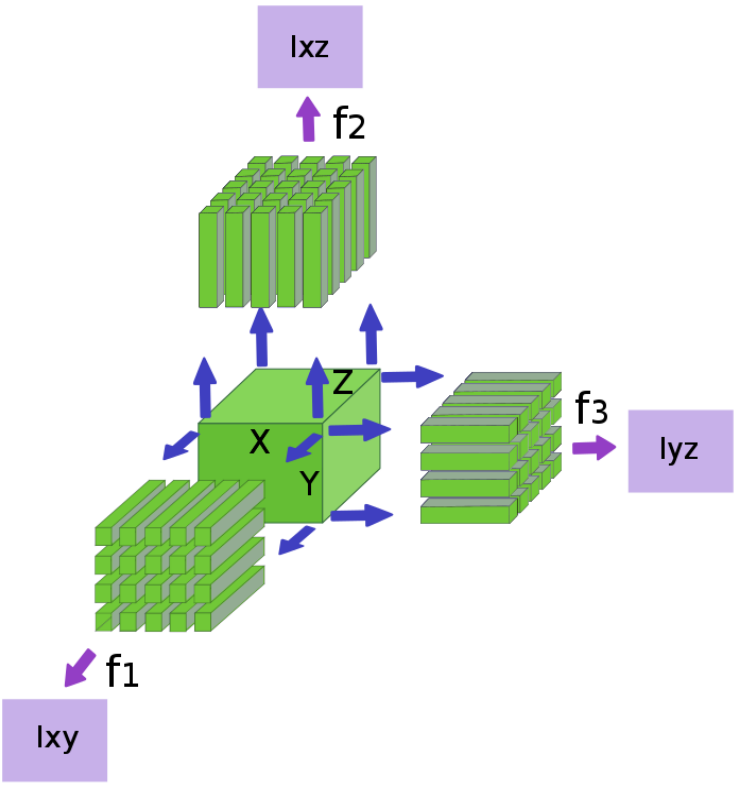


Figura 8. Seqüència projectada en 3 vistes simples.

Per calcular les projeccions I_{xy}, I_{xz}, i I_{yz}, partim de les components del *Optical flow* (OF) o del RGB de les seqüències de vídeo. Les projeccions presentades preserven informació molt útil per al reconeixement d'accions humanes i condensen informació de tota una seqüència de vídeo en tres components. I_{xy} proporciona informació de la relació de la magnitud de les parts en moviment. I_{xz} proporciona informació de la relació de la magnitud de moviment horitzontal i I_{yz} proporciona informació de la relació de la magnitud de moviment vertical. Quan les projeccions han estat calculades partint de les components F_x i F_y, aquestes també proporcionen informació de la direcció del moviment. Per tant, I_{xz} i I_{yz} retenen informació espai-temporal sobre el moviment dels objectes, mentre que I_{xy} reté informació de la aparença del moviment. La figures 9, 10 i 11 mostren exemples de les 3 projeccions aplicades a les accions *bend*, *run* i *jack* del *dataset* Weizmann utilitzant la funció *Max* a f_1, f_2, i f_3.

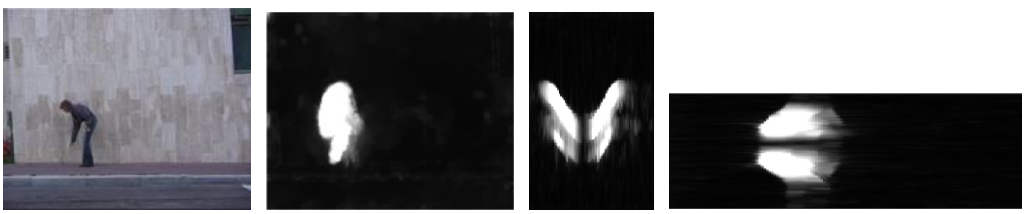


Figura 9. Mostra d'un *frame* i de les vistes I_{xy}, I_{yz} I_{xz} per l'acció *bend* del Weizman data set utilitzant f_1=*max*, f_2=*max* and f_3=*max*.

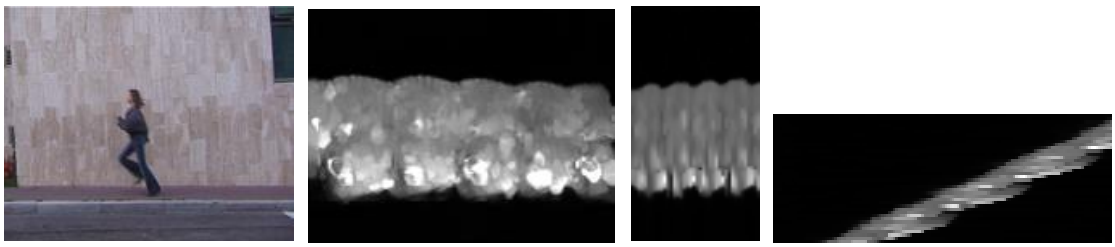


Figura 10. Mostra d'un *frame* i de les vistes I_{xy}, I_{yz} I_{xz} per l'acció *run* del Weizman *dataset* utilitzant f_1=*max*, f_2=*max* and f_3=*max*.

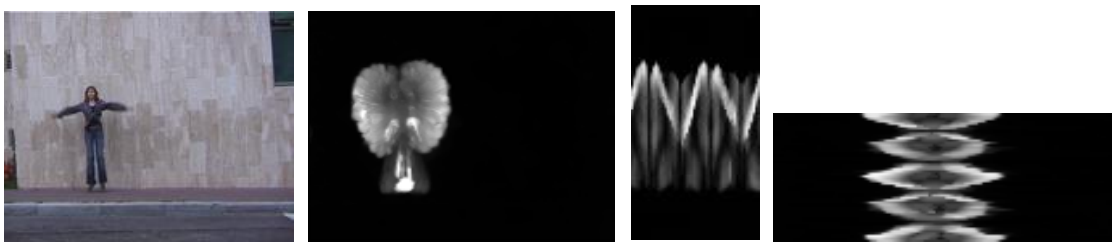


Figura 11. Mostra d'un *frame* i de les vistes I_{xy}, I_{yz} I_{xz} per l'acció *jack* del Weizman *dataset* utilitzant f_1=*max*, f_2=*max* and f_3=*max*.

3.3 Projecció del tensor

Ja que la projecció d'un triplet (I_{xy}, I_{xz} i I_{yz}) condensa informació d'una seqüència a un única imatge, la projecció I_{xy} pot perdre informació important a seqüències molt llargues, especialment a seqüències de vídeo que contenen més d'una acció. Per minimitzat aquest inconvenient, hem experimentat amb dos diferents escenaris.

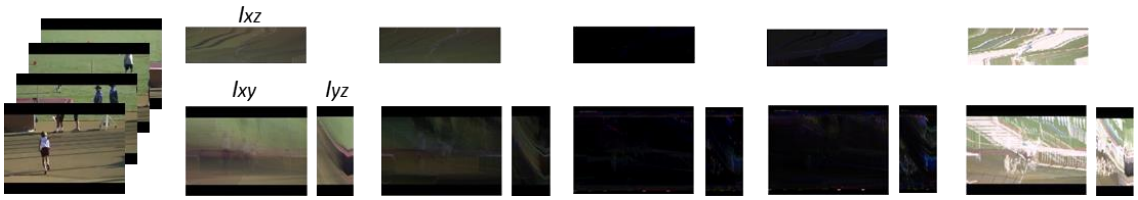


Figura 12. D'esquerra a dreta, (a) mostra de l'acció *HighJump* del *dataset* UCF101, (b) Projecció fent servir *f=Mean*, (c) Projecció fent servir *f=StDev*, (d) Projecció fent servir *f=Skewness*, (e) Projecció fent servir *f=Kurtosis* i (f) Projecció fent servir *f=Max*.

3.3.1 Projecció de vista simple

A un primer escenari hem calculat una única imatge des de cada projecció. Per tant, tenim un triplet I_{xy}, I_{xz} i I_{yz} per cada vídeo seqüència. La figura 12 mostra aquest escenari per a una vídeo seqüència RGB. D'ara endavant, utilitzarem *S* per indicar que les projeccions van ser calculades utilitzant vista simple.

3.3.2 Projecció de vista múltiple

A un segon escenari, com el suggerit a [78], hem generat múltiples *templates* des de cada seqüència de vídeo trencant-les en múltiples segments sobreposats. Hem utilitzat unes finestres temporal de mida τ sobreposades ε *frames*. El principal inconvenient d'aquest escenari, és que també s'incrementa la mida del *dataset*. D'ara endavant, utilitzarem *M* per indicar que les projeccions van ser calculades utilitzant múltiples segments sobreposats.

3.4 Estudi del problema de l'auto-oclusió

Encara que el triplet de projeccions no pot solucionar completament el problema de l'auto-oclusió, aquest pot ser reduït de forma important. L'èxit de les projeccions per solucionar aquest problema depèn de la funció utilitzada per calcular les vistes I_{xy}, I_{xz} i I_{yz}. En aquest apartat s'ha estudiat el problema de l'auto-oclusió utilitzant la funció *Max*, i s'ha observat que aquesta és reduïda significativament. Això és degut al fet que les *templates* són calculades a partir de la projecció dels 3 modes de les fibres de la seqüència d'imatges, i per tant, alguns píxels poden ser representats en més d'una vista.

I_{xy}, I_{xz} i I_{yz} han estat calculades a partir de les fibres de mode3, fibres de mode2 i fibres de mode1 respectivament. Per tant, cada fila a la vista I_{xz} i cada columna a la vista I_{yz} representa un *frame* de la seqüència de vídeo. Les figures 13 i 14 mostren aquestes vistes per l'acció *run* del Weizmann *data set.*

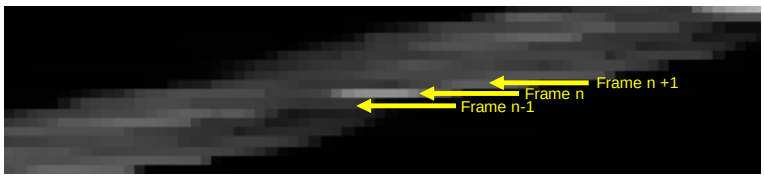


Figura 13. Vista magnificada de I_{xz} calculada amb f_2=*max*. Cada fila correspon a un *frame* de la seqüència.

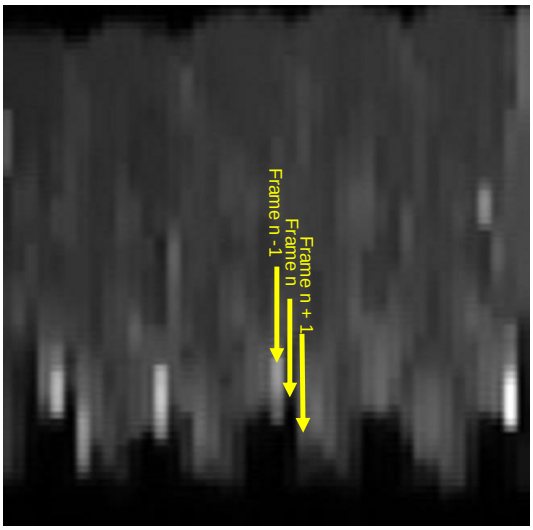


Figura 14. Vista magnificada de I_{yz} calculada amb $f_3=max$. Cada columna correspon a un *frame* de la seqüència.

Per estudiar el problema de l'auto-oclusió, s'ha analitzat el número de vegades que cada píxel *x*, *y* queda representat en els triplets. És obvi que la vista I_{xy} només pot representar cada píxel una sola vegada, en canvi, les vistes I_{xz} i I_{yz} poden representar el mateix píxel moltes vegades, corresponents a diferents *frames* de la seqüència de vídeo. La figura 15 mostra una parella d'imatges de les acumulacions, on el color indica quantes vegades queda representat cada píxel en els triplets.

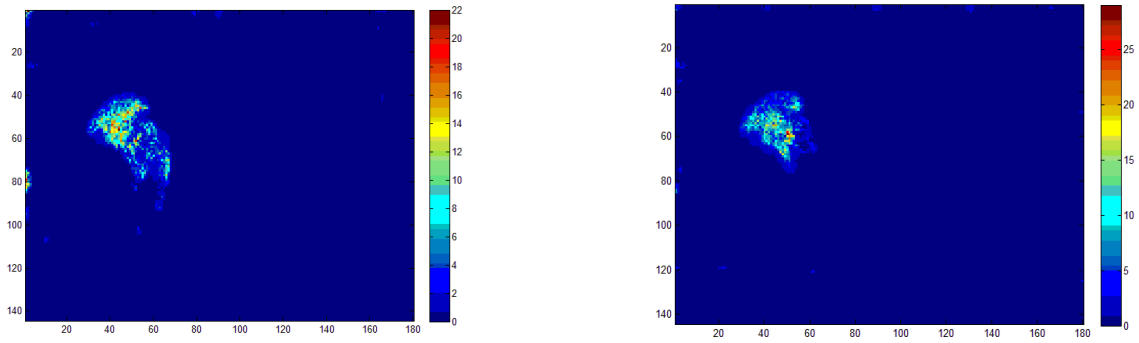


Figura 15. Número de vegades que cada píxel *x*,*y* es representat en el triplet per a l'acció *wave1* del *dataset* Weizmann. **Esquerra:** F_x component. **Dreta:** F_y component.

La figura 15 mostra que alguns píxels a la regió de moviment són seleccionats moltes vegades en els triplets utilitzant la funció *Max*. Alguns píxels a la component F_x són seleccionats fins a 20 vegades, mentre que a la component F_y alguns píxels són seleccionats fins a 25 vegades, per tant, el valor d'un píxel donat pot ser representat a les projeccions amb molts diferents valors corresponent al seu valor a diferents *frames*.

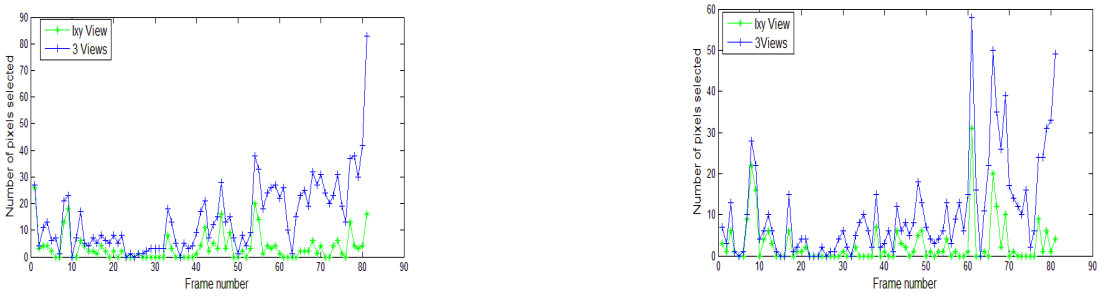


Figura 16. Número de píxels seleccionats per un triplet i la vista I_{xy} de cada *frame* de una vídeo seqüència. **Esquerra:** F_x component. **Dreta:** F_y component.

A la figura 16, s'ha comparat el número de píxels seleccionats a cada *frame* de la seqüència de vídeo en els triplets i els seleccionat per la vista I_{xy} únicament. Aquesta figura mostra que treballant amb les 3 projeccions, es pot retenir informació d'un major número de píxels que amb la vista I_{xy} únicament.

Tot i que les figures 15 i 16 mostren que les tres projeccions retenen més informació que una única vista com la I_{xy}, això no prova que el problema de l'auto-oclusió quedi reduït, perquè no hem provat que els píxels retinguts múltiples vegades hagin estat retinguts amb informació de diferents direccions a les diferents retencions. La figura 17 mostra l'historia de les direccions de 3 píxels diferents retingudes al triplet i les direccions retingudes utilitzant únicament la vista I_{xy} per l'acció *wave1* del Weizmann *dataset* que conté auto-oclusió.

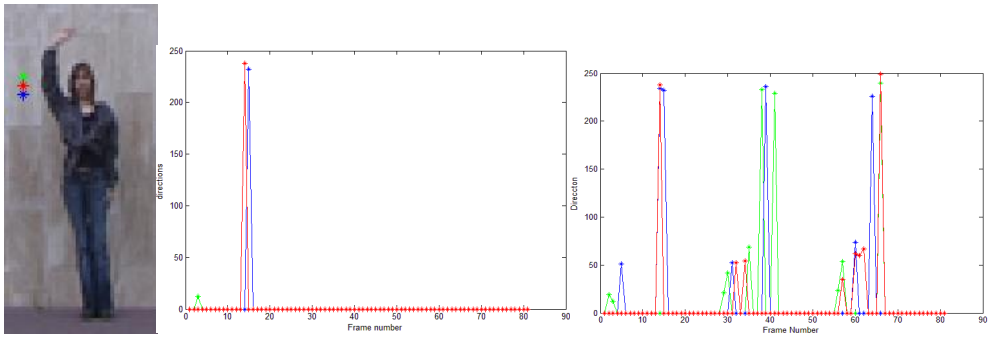


Figura 17. Historia de les direccions per 3 píxels diferents a una regió de moviment per a l'acció *wave1* del *dataset* Weizmann. Les direccions corresponen a 0-360 graus. **Esquerra**: Imatge dels píxels seleccionats. **Centre**: Historia dels píxels per a la vista I_{xy}. **Dreta**: Historia dels píxels per a les tres vistes.

Aquesta figura mostra que la vista I_{xy} només reté una única direcció per a cada píxel, mentre que les 3 projeccions considerades conjuntament, retenen les dos direccions principals corresponents als moviments de pujada i baixada.

Les figures 19 i 20 mostren un altre exemple de la informació retinguda per les vistes I_{xz} i I_{yz} per l'acció *jack* del *dataset* Weizmann, que conté moviment de braços i cames simultàniament. La figura 18 mostra un exemple dels moviments que conté aquesta seqüència.

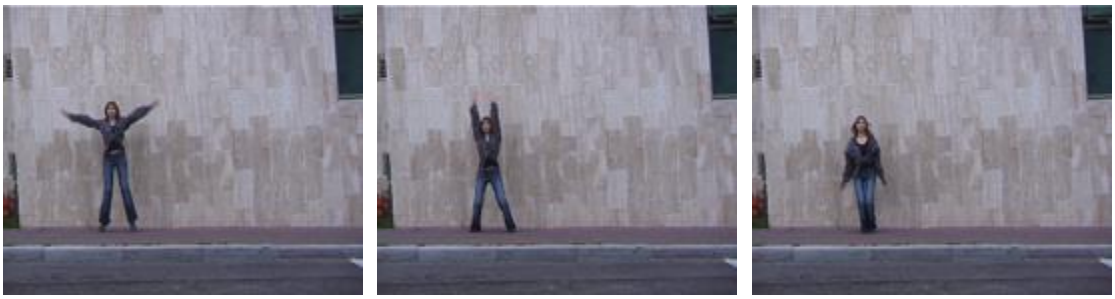


Figura 18. Mostra de l'acció *jack* del *dataset* Weizmann.

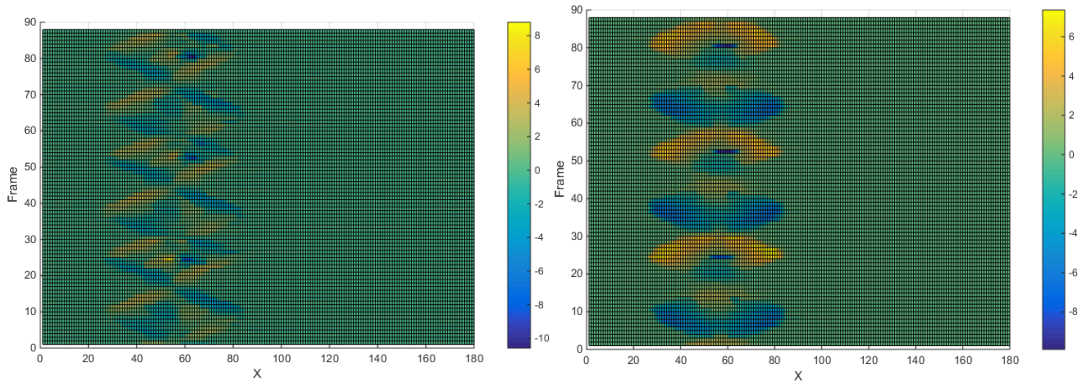


Figura 19. Mostra de la Vista I_{xz} per l'acció *jack* del *dataset* Weizmann. **Esquerra**: component F_x. **Dreta**: component F_y.

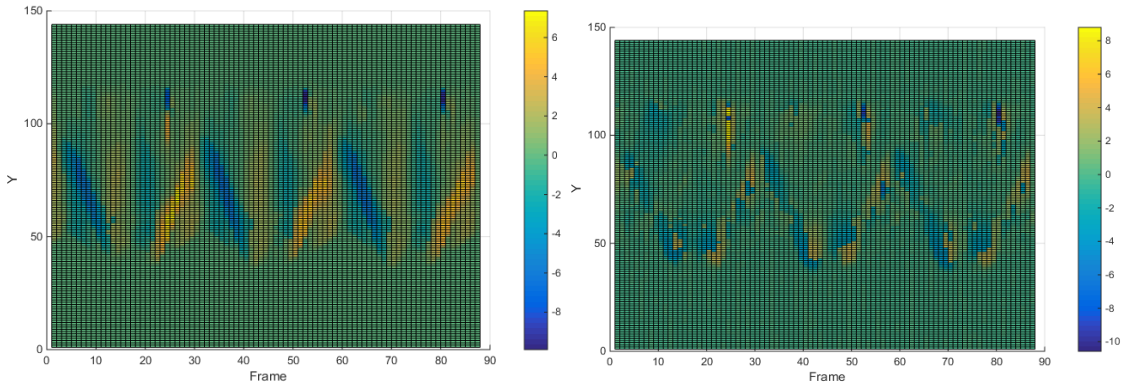


Figura 20. Mostra de la Vista I_{yz} per l'acció *jack* del *dataset* Weizmann. **Esquerra**: component F_x. **Dreta**: component F_y.

Les figures 19 i 20 mostren com les vistes I_{xz} i I_{yz}, conserven informació de direcció de les diferents parts del cos. La figura 19 dreta, mostra que la vista I_{xz} reté informació de la pujada i baixa del moviment dels braços, la figura 19 esquerra mostra que I_{xz} reté informació del moviment dels braços d'esquerra i dreta.

La figura 20 esquerra, mostra que la vista I_{yz} reté informació del moviment a dreta i esquerra de les cames, la figura 20 dreta mostra la informació retinguda de pujada i baixada del moviment de les cames. Es pot veure que I_{xz} reté informació del moviment de braços i I_{yz} informació del moviment de les cames principalment. Per tant, les dues vistes són útils, ja que cadascuna descriu una regió diferent de la seqüència d'imatges.

Tot això prova que considerant les tres projeccions conjuntament, no es sobreescriu completament la informació prèvia, i que per tant, això redueix el problema de l'auto-oclusió. A més, tot i que les accions considerades en aquest treball no contenen diferents velocitats en una mateixa acció, està clar que considerant les tres projeccions, podríem retenir informació de diferents velocitats.

3.5 Extracció/descripció de característiques

Ja que I_{xy}, I_{xz} i I_{yz} són imatges 2D, poden ser utilitzats els algorismes estàndard de detecció i descripció de *keypoints* d'imatges de nivell de gris. Els vectors de característiques són utilitzats per descriure el veïnat dels *keypoints* de la imatge. Aquests *keypoints*, poden ser extrets utilitzant un detector estàndard de *keypoints* (ex. SIFT [17]) o per mostreig dens de la imatge (ex. PHOW [20]). Les coordenades espacials dels *keypoints* obtinguts van acompanyades per les seves escales, les quals defineixen la extensió del veïnat. El contingut de les imatges projectades (I_{xy}, I_{xz}, I_{yz}) és representat com un conjunt de descriptors corresponents al *keypoints* obtinguts. Aquets descriptors, extrets de les imatges d'entrenament, són utilitzats per identificar *keypoints* similars en noves imatges projectades des de noves seqüències.

La figura 21 mostra els *keypoints* obtinguts utilitzant l'algorisme PHOW. Ja que PHOW calcula una gran quantitat de *keypoints*, i amb la intenció de millorar la visualització, a la figura només mostrem 50 *keypoints* seleccionats aleatòriament. Els cercles estan centrats a la localització dels *keypoints*, els seus radis són les escales, i les línies de dintre són les orientacions principals. Es pot veure que molts *keypoints* estan localitzats a regions on les accions van ser realitzades i les seves orientacions són molt similars entre elles. Una vegada els *keypoints* han estat localitzats, s'aplica SIFT en la seva forma tradicional a cada *keypoint*, ja que PHOW és simplement un SIFT dens en diferents resolucions.

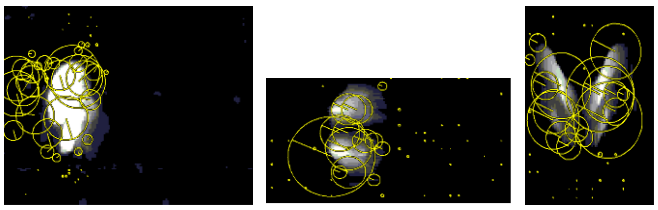


Figura 21. Mostra de PHOW aplicat sobre el triplet calculat per a l'acció *bend* (a) I_{xy} projecció, (b) I_{xz} projecció (c) I_{yz} projecció.

Uns altres descriptors de característiques com HOG [18], poden ser aplicats a la imatge global sense una prèvia detecció de *keypoints*. Les imatges de gradients locals d'intensitat, o les direccions de les transicions, són dividides en petites regions anomenades cells i acumulades en un histograma 1D. A la següent etapa, aquests histogrames alimentaran el classificador.

La figura 22 mostra les característiques HOG obtingudes a I_{xy} , I_{xz} i I_{yz}. Es pot veure que una gran quantitat d'informació de gradients està localitzada a regions on les accions van ser realitzades.

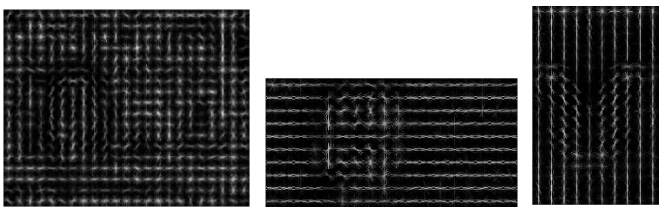


Figura 22. Mostra de HOG aplicat sobre el triplet calculat per a l'acció *bend* (a) I_{xy} projecció, (b) I_{xz} projecció (c) I_{yz} projecció.

El descriptor LIOP també pot ser aplicat a les imatge 2D. LIOP pot ser combinat amb diferents detectors de característiques com *Difference Of Gaussians* (DOG) [17], Harris-Laplace [64], etc. La figura 23 mostra les característiques DOG obtingudes a I_{xy} , I_{xz} i I_{yz} . Encara que la figura mostra que els *keypoints* estan localitzats a les regions on s'han produït les accions, aquest *keypoints* són molt sensibles als paràmetres de sintonia de DOG (número d'octaves, número de nivells per octava, llindar de pics, etc.). Una vegada localitzats els *keypoints*, s'aplica el descriptor LIOP a cada un d'ells.

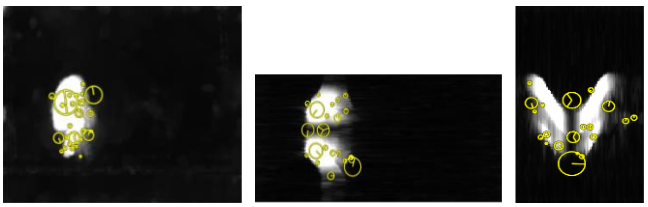


Figura 23. Mostra de LIOP aplicat sobre el triplet calculat per a l'acció *bend* (a) I_{xy} projecció, (b) I_{xz} projecció (c) I_{yz} projecció.

La figura 24 il·lustra les característiques *C1* extretes a la fase d'entrenament per SMFs. Cada projecció es representada per quatre imatges corresponents a quatre diferents orientacions dels filtres *Gabor*. Es pot veure la diferent resposta de cada filtre per a la mateixa regió de la imatge.

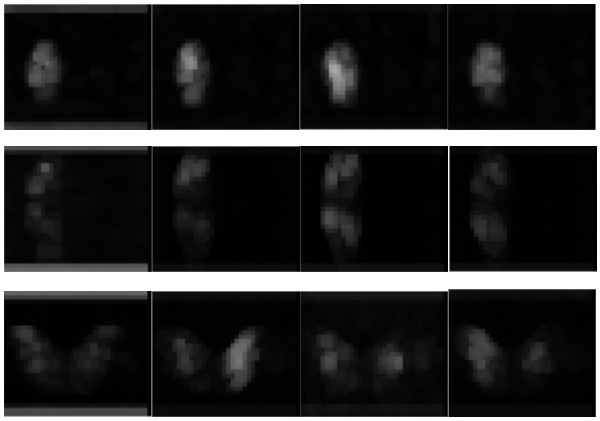


Figura 24. Mostra de les característiques *C1* SMFs per a l'acció *bend*. Les files de dalt a baix son les projeccions I_{xy} , I_{xz} i I_{yz} respectivament. Les columnes corresponen d'esquerra a dreta a les orientacions -45°, 0°, 90° i 45° dels filtres *Gabor*.

3.6 Representació de característiques

Per representar les característiques resultants de les *templates*, hem utilitzat dos tècniques diferents anomenades *Bag of Features* (BoF) i *Fisher Vector Encoding* (*Fv*). La aproximació BoF es caracteritza per l'ús d'una col·lecció desordenada de característiques de la imatge, perdent qualsevol informació d'estructura o espacial. Així, els descriptors similar són agrupats utilitzant l'algorisme *k-means*. Els centres d'aquest grups (*clusters*) defineixen un *Visual Codebook*. El número de *clusters* és la mida del *Visual Codebook* i cada *cluster* al *Visual Codebook* forma una paraula visual. Una vegada el diccionari ha estat construït, es calcula un histograma per cada triplet a l'aprenentatge. Cada histograma modela una acció humana. Després construïm un *randomized K-D tree forest* [65] de paraules visuals per accelerar l'execució.

A [44], van provar que *Fv* supera a BoF en reconeixement d'accions humanes. Els *Fv* estenen BoF però codifiquen els moment estadístics de primer i segon ordre dels descriptors de característiques i un *Gaussian Mixture Model* (GMM). *Fv* descriu la desviació d'un conjunt de descriptors des d'una distribució mitjana de descriptors calculada a partir d'un model paramètric generatiu. Hem estudiat *Fv* per avaluar la millora de rendiment aconseguida a les nostres *templates*. Hem reduït la dimensió dels descriptors fent servir *Principal Component Analysis* (PCA) i aplicant normalització $\ell 2$ per a *Fv*.

3.7 Classificació d'accions

A la fase de classificació el classificador s'ha alimentat amb la sortida dels descriptors de característiques obtinguts. S'han fet quatre implementacions diferents depenent del extractor/descriptor de característiques (PHOW, HOG, SMFs i LIOP) utilitzat i una darrera implementació d'alt rendiment que combina més d'una funció de projecció, *Fv* i fusió de les característiques amb les de IDTs. A cada implementació s'han fet servir els mateixos classificadors reportats pels seus autors en els papers originals tant per a PHOW, HOG, i SMFs [10, 18 i 8]. En el cas de LIOP, ja que els seus autors no el presenten amb cap classificador concret, s'ha utilitzat un model combinat de BoF i SVM.

Per tant, s'han combinat PHOW i LIOP amb BoF i un classificador SVM a HOG i SMFs, s'ha aplicat un classificador SVM directament per evitar la dependència de una sola tècnica d'extracció/descripció de característiques i de un sol tipus de classificador.

La figura 25 mostra les quatre combinacions utilitzades.

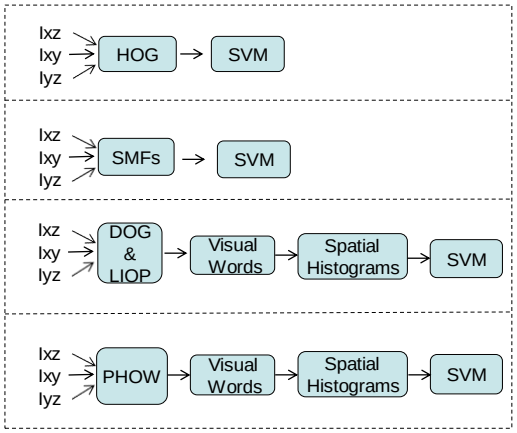


Figura 25. Diagrama de blocs de les quatre implementacions utilitzades per avaluar el nostre sistema.

Les figures 26, 27, 28 i 29 mostren més detalls de les implementacions utilitzades.

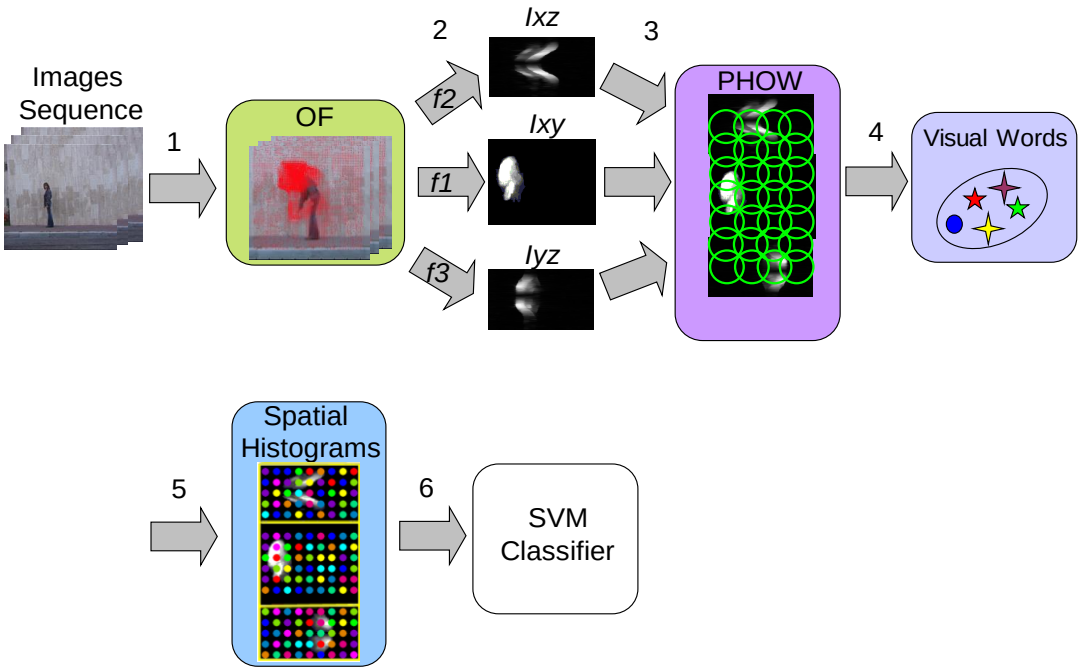


Figura 26. Diagrama de blocs complet del procés fent servir PHOW.

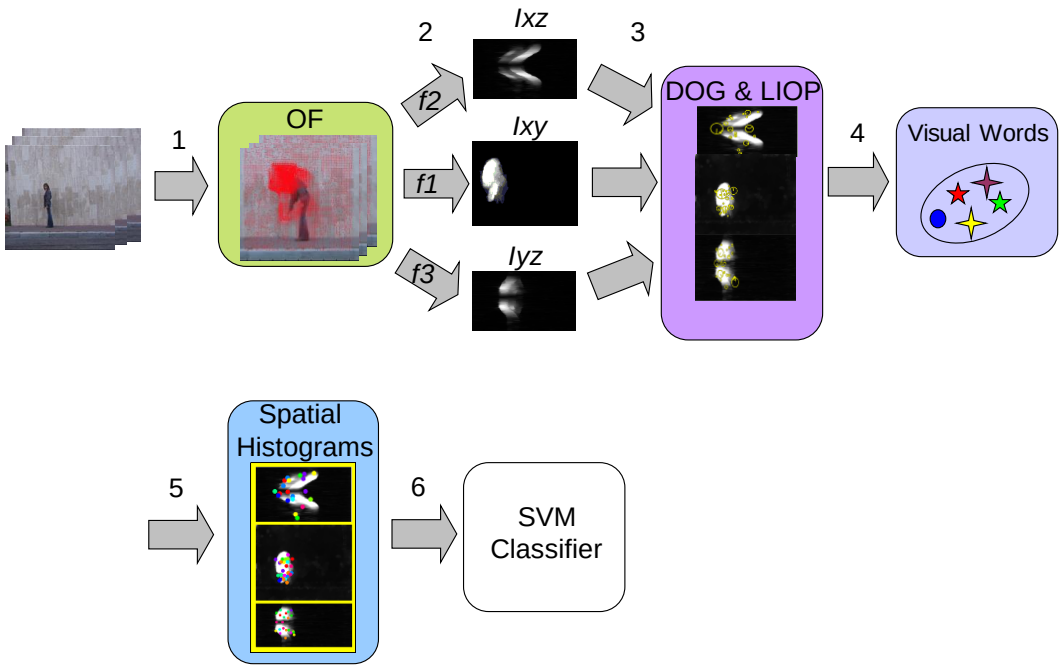


Figura 27. Diagrama de blocs complet del procés fent servir DOG&LIOP.

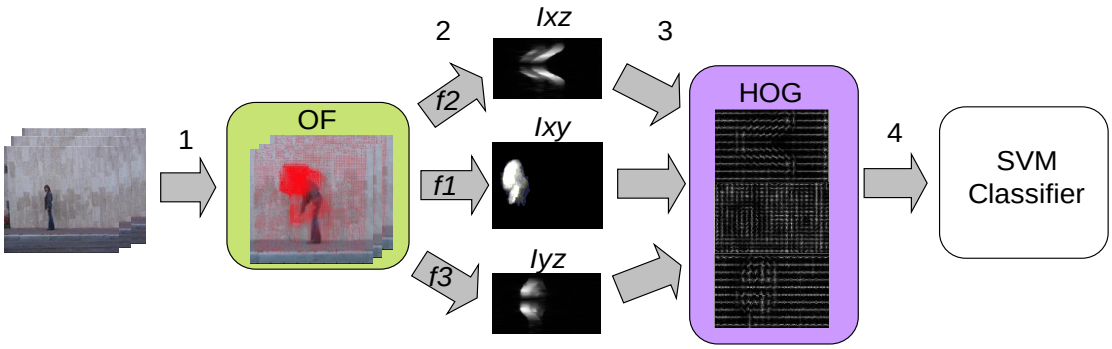


Figura 28. Diagrama de blocs complet del procés fent servir HOG.

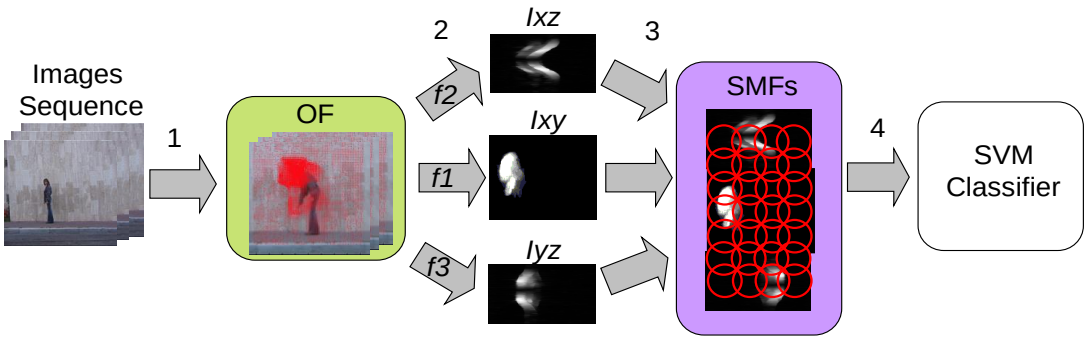


Figura 29. Diagrama de blocs complet del procés fent servir SMFs.

S'ha entrenat el sistema per construir un vocabulari visual agrupant les característiques PHOW/LIOP extretes de les imatges d'entrenament. Després, les característiques de les imatges d'entrenament són detectades i mapejades als seus termes més propers del vocabulari visual. El resultat és un histograma de característiques quantificades detectades a les imatges d'entrenament. Finalment, aquests histogrames són utilitzats per entrenar una SVM lineal. A la fase de reconeixement, les característiques de les imatges de test són detectades i mapejades als seus termes més propers del vocabulari visual i els histogrames resultats alimenten el classificador SVM.

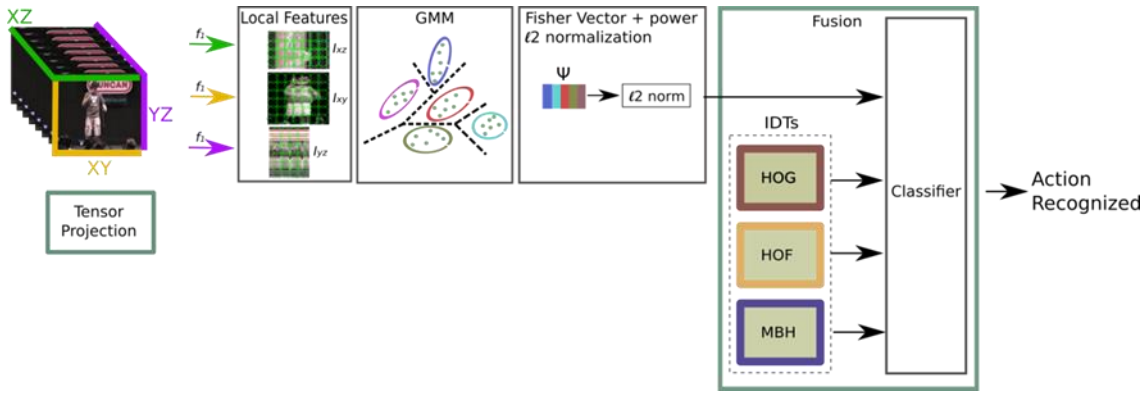


Figura 30. Exemple de la implementació completa del mètode.

Les implementacions anteriors, només utilitzen una funció de projecció per calcular les *templates*. A la implementació completa del mètode, es calculen les *templates* per a les funcions *Max*, *Mean*, *Standard Deviation*, *Kurtosi* i *Skewness*, es calculen els *Fv* i es fusionen les característiques entre elles. Finalment, les característiques resultants es fusionen amb les característiques del HOG, HOF i MBH (IDTs).

La figura 30 mostra un exemple d'aquesta darrera implementació.

4 Resultats

4.1 Objectius dels experiments

Per fer més clar el gran número d'experiments realitzats, hem separat aquests en dos grups. Al primer grup d'experiment s'estudia el rendiment de les *templates* per classificar accions humanes a *datasets* simples. Al segon grup s'estudia el rendiment de les *templates* per classificar accions a entorns més complexos utilitzant uns *datasets* de més complexitat.

4.2 Datasets

Pel desenvolupament d'aquesta tesi s'han utilitzat els següents sis *datasets* públics: Weizmann [93], KTH [94], Cambridge Hand-Gesture [107], UCF101 [95], HMDB51 [96] i Hollywood [106].

Els *datasets* Weizmann, KTH i Cambridge Hand-Gesture són simples, ja que a les imatges només apareix la persona o la mà que realitza l'acció. Els *datasets* UCF101, HMDB51 i Hollywood en canvi són *datasets* complexos d'accions de la vida real on poden aparèixer diferents persones en cada seqüència. Aquets *datasets* han estat àmpliament utilitzats en reconeixement d'accions i gestos humans.

En els següents sis capítols es dóna una descripció més detallada de cadascun dels *datasets*.

4.2.1 Dataset Weizmann

Aquest *dataset* està format per 10 accions diferents fetes per 9 persones diferents. Ha estat gravat amb càmera i *background* estàtics; no existeixen oclusions i només hi ha una única persona en moviment en cada seqüència. No presenta canvis importants d'il·luminació. La figura 31 mostra alguns exemples d'aquestes accions. En aquest treball s'ha utilitzat el mètode *leave-one-out cross validation* per la fase de test, ja que aquest és el mètode normalment utilitzat pels altres autors per testejar amb aquest *dataset*. 10 accions fetes per una persona s'utilitzen per test i les restants 8 persones són utilitzades per entrenament. Aquest procés es repeteix per cada una de les 9 persones. Per tant, al final del test tindrem el *Recognition Rate* (RR) de cada una de les 9 persones per a cada una de les 10 accions.

Figura 31. Weizmann human actions. *Bend, jack, skip, jump, run, side, skip, walk, wave1, wave2*.

4.2.2 Dataset KTH

Aquest *dataset* està format per 6 tipus d'accions humanes fetes per 25 persones. Hi ha una única persona movent-se a cada seqüència, no existeixen oclusions però existeixen variacions d'escala, i gent amb roba diferent. Conté seqüències gravades a exteriors i seqüències gravades a interiors. A la figura 32 es mostren exemples d'algunes accions d'aquest *dataset*. A aquest *dataset* n'hi ha 600 seqüències de vídeo.

Com al *dataset* anterior, aquí s'ha fet servir el mètode *leave-one-out cross validation* per test. Sis accions fetes per una persona són utilitzades per test i les restants 24 persones són utilitzades per entrenament. Aquest procés es repeteix per cada una de les 25 persones, i al final tenim el RR per a cada persona i cada acció.

Figura 32. KTH human actions. *Boxing, handclapping, handwaving, jogging, running, walking*.

4.2.3 Dataset Cambridge Hand-Gesture

El *dataset* Cambridge hand-gesture està format per 900 seqüències de vídeo amb 9 gestos de mans diferents (100 seqüències de vídeo per classe de gest). Ha estat gravat en cinc il·luminacions diferents i els gestos han estat agrupats per diferents conjunts d'il·luminació. La figura 33 mostra alguns exemples. Seguint el protocol experimental de [16], el *dataset* es trenca en un número de conjunts d'il·luminació on Set1, Set2, Set3 i Set4 són utilitzats per test i el Set5 per entrenament. El Set5, a més, és trencat en un conjunt d'entrenament i un altre conjunt de validació (90 seqüències de vídeo per l'entrenament i 90 seqüències de vídeo per validació). S'utilitzen cinc subconjunts aleatoris per entrenament i validació del conjunt Set5.

Figura 33. Cambridge hand-gesture *samples*. *Flat-Leftward, Flat-Rightward, Flat-Contract, Spread-Leftward, Spread-Rightward, Spread-Contract, V-Shape-Leftward, V-Shape -Rightward and VShape-Contract*.

4.2.4 Dataset UCF101

Aquest *dataset* es compon per 101 classes d'accions humanes. Té 13.320 clips de vídeo amb *frames* de 320 x 240 píxels de resolució. Hem dut a terme l'avaluació d'acord amb les tres divisions d'entrenament /test entregats amb aquest *dataset*. Cada vídeo conté una sola acció en diferents entorns i *background* poc massificat. La figura 34 mostra un exemple del *dataset* UCF101[95].

Figura 34. UCF101 human actions.

4.2.5 Dataset HMDB51

Es tracta d'un *dataset* molt complicat per al reconeixement d'accions humanes. Conté 51 accions humanes en 6766 vídeos. Els vídeos contenen una sola acció en escenaris realistes descarregats de YouTube. Hi ha canvis d'escala i d'il·luminació. Hem dut a terme l'avaluació d'acord amb les tres divisions d'entrenament /test entregats amb aquest conjunt de dades. La figura 35 mostra un exemple del *dataset* HMDB51[96].

Figura 35. HMDB51 human actions.

4.2.6 Dataset Hollywood

Aquest *dataset* consisteix en 430 vídeos incloent-hi seqüències curtes de 32 pel·lícules. La resolució varia de 300 x 200 a 400 x 300 píxels depenen dels vídeos. Ha estat gravat amb càmera no-estàtica, les seqüències contenen *background* desordenat i oclusions entre persones. La figura 36 mostra algunes accions d'aquest *dataset*.

Figura 36. Hollywood human actions.

4.3 Experiments amb datasets simples

L'objectiu principal d'aquest experiment és sintonitzar la tècnica proposada amb els paràmetres òptims per aconseguir el millor rendiment en reconeixement d'accions humanes. Els paràmetres que han de ser determinats són: les funcions f_i utilitzades per projectar els subtensors, les amplades dels tractes β i els detector/descriptor de *keypoints* a utilitzar.

Amb aquest objectiu, s'han calculat múltiples triplets de (I_{xy} , I_{xz} , I_{yz}) utilitzant cinc funcions (*supremum, Mean, Standard Deviation, Kurtosi* i *Skewness*), amb sis valor diferents de veïnat (β =0, 2, 4, 6, 8 i 10). Per tant, tindrem 30 diferents triplets. S'han fixat les tres funcions de projeccions iguals a tots els experiment, es a dir, set $f_1 = f_2 = f_3$.

Després, s'han aplicat quatre detectors/descriptor de l'estat de l'art (PHOW, LIOP, HOG i SMFs) a tots els triplets obtinguts. En aquesta part de l'experiment s'avalua quina tècnica d'extracció/descripció de característiques és més adequada per a la nostra aproximació.

En un segon experiment, es mostren els resultats obtinguts per l'aproximació presentada respecte al MHI clàssic. Per fer això s'ha aplicat sobre la imatge MHI els detectors/descriptor (PHOW, LIOP, HOG i SMFs). Amb els *keypoints* obtinguts s'ha intentat reconèixer accions humanes de les seqüències dels *datasets*, i s'han comparat els resultats amb els prèviament obtinguts a la nostra aproximació. L'objectiu d'aquest experiment és avaluar el nostre triplet de projeccions respecte al *template* generat per MHI, així com l'efecte de diferents funcions de projecció f_i.

Una vegada s'ha determinat la funció òptima de projecció f_i, la mida del tracte β , i el descriptor de *keypoints*, s'ha fet un tercer experiment amb l'objectiu de comparar l'aproximació aquí presentada respecte a diferents tècniques de l'estat de l'art. Hem comparat els resultats obtinguts utilitzant diferents *datasets* públics utilitzats normalment com a referència.

4.4 Experiments amb datasets complexos

L'objectiu d'aquest grup d'experiments és mostrar que el mètode presentat basat en la projecció de les seqüències és complementari amb IDTs i que per tant, aquest és útil per millorar el rendiment d'aquest últim en el reconeixement d'accions humanes. Per provar això, comparem els resultat obtinguts, amb els publicats per altres autors a la literatura recent utilitzant els mateixos *datasets*.

En un primer experiment hem avaluat si la representació de característiques *Fv* millora el rendiment respecte a BoW.

Al segon experiment hem comparat el rendiment utilitzant un únic o múltiples projeccions dels tensors. $S_{function}$ computa una única imatge per cada projecció, mentre que $M_{function}$ genera múltiples *templates* per cada vídeo seqüència trencant aquests en múltiples segments sobreposats.

A un tercer experiment hem estudiat la complementarietat de les cinc funcions de projecció *Max, Mean Standard Deviation, Skewness* i *Kurtosi* mesclant-les fent servir *Sum Pooling*, i avaluant la millora de rendiment aconseguida quan afegim cada una d'elles al classificador.

Una vegada tots els paràmetres han estat fixats, hem afegit el nostre mètode a IDTs i avaluat la millora de rendiment proporcionada per les nostres *templates*. Finalment, hem comparat la nostra aproximació amb altres tècniques de l'estat de art a dos *datasets* públics. Amb aquest propòsit, hem fixat tots els paràmetres als valors que van donar el millor rendiment als experiments previs.

4.5 Criteri d'avaluació

Amb l'objectiu de mesurar el rendiment de la nostre tècnica, una vegada ha estat entrenat, mesurarem el *Recognition Rate* ($\overline{RR}$) de les seqüències de test utilitzant la següent equació:

$$\overline{RR} = \frac{\#samples \quad correctly \quad classified}{Total \quad samples \quad Tested} \; . \qquad (10)$$

Per a cada funció de projecció (*Max, Mean, StDev, Skewness i Kurtosi*), cada veïnat β, i cada extractor/descriptor de *keypoints* (PHOW, LIOP, HOG i SMFs), es calcula el $\overline{RR}$.

Fent servir dos *datasets* d'accions humanes simples, s'ha calculat el $\overline{RR}$ mig d'una funció de projecció donada pels quatre extractor/descriptor de *keypoints* testejats. També s'ha calculat el $\overline{RR}$ mig de un extractor/descriptor de *keypoints* donat per a totes les funcions de projecció.

Una vegada hem determinat la funció de projecció i el veïnat òptim, aquests valors han estat utilitzats als següents experiments. S'han aplicat novament els mateixos extractor/descriptor de *keypoints* a les imatges MHI, i comparat els seu rendiment contra les nostres projeccions. Considerarem el rendiment mig dels quatre extractor/descriptor de *keypoints* i el millor obtingut per MHI i els nostres triplets per avaluar els nostres resultats.

Després, comparem la nostra aproximació amb altres tècniques de l'estat de l'art en els dos *datasets* públics de reconeixement d'accions més simples. Amb aquest propòsit, s'ha fixat la combinació de la tècnica d'extracció/descripció, funció de projecció i valor β que han obtingut el millor al primer experiment per fer la comparativa. S'han considerat els resultats publicats pels autors per fer la comparativa. Fent servir un *dataset* de gestos humans, hem mesurat també els resultats que s'obtenen amb la tècnica aplicada al reconeixement de gestos sobre un altre *dataset* públic.

Fent servir dos *datasets* més complexos, hem avaluat quina és la millor representació de característiques per al nostre sistema *Fv* o BoW (per a cada projecció individualment i per les tres combinades), i quin dels dos escenaris, $S_{function}$ o $M_{function}$, obté millor rendiment. També hem estudiat la complementarietat de les diferents funcions (*Max*, *Mean*, *StDev*, *Skewness i Kurtosi*), mesurant els rendiment aportat per cadascuna d'elles quan són afegides utilitzant l'algorisme *Sequential Forward Selection* (SFS) [66]. Finalment, hem fusionat les nostres característiques amb les de IDTs i comparat els resultats obtinguts amb els de l'estat de l'art en aquest dos *datasets* més complexos.

4.6 Resultats experimentals per datasets simples

En aquesta secció, presentem i discutim els resultats experimentals de l'avaluació. Tots els experiments han estat testejats en els *datasets* Weizmann i KTH. Mostrarem els millors resultats o els resultats mitjans obtinguts per a cada funció de projecció i per a cada tècnica d'extracció/descripció.

Per a PHOW s'ha utilitzat *step* = 5, *size*=7, on *step* és la distància entre els *keypoints* i *size* és la mida del descriptor. Per a PHOW i LIOP s'han testejat *codewords* des de 100 a 1100 paraules visuals. Per HOG s'ha testejat des de 9 a 36 *bins* i *cells* variables des de 3 a 15 píxels. S'han utilitzat 1000 *patches* i 4 orientacions als filtres (-45º, 0º, 90º, 45º).

4.6.1 Test de funcions de projecció en tractes

La taules 1, 2, 3, 4, 5 i 6 mostren els resultats obtinguts al *dataset* Weizmann utilitzant les funcions de projecció *Max*, *Mean*, *Skewness*, *Kurtosi* i *StDev* per un veïnat β =0, 2, 4, 6, 8 i 10 . Les columnes mostren el resultats obtinguts per cada tècnica d'extracció/descripció de característiques.

Les files mostren els resultats obtinguts utilitzant les projeccions generades per cada diferent funció. La darrera fila i columna mostren el rendiment mig.

	$\beta = 0$				
Method	PHOW	LIOP	HOG	SMFs	Avg
Max	96.6	92.2	82.2	85	89
Mean	95.5	91.1	91.1	87.7	91.4
Skew	92.2	73.3	75.5	83.3	81.1
Kurt	90.1	78.1	77.7	74.1	80
StDev	**98.8**	95.5	88.8	92.2	93.8
Avg	95.8	88	84.4	87	

Taula 1. *Recognition Rate* (%) per al *dataset* Weizmann per a $\beta = 0$.

	$\beta = 2$				
Method	PHOW	LIOP	HOG	SMFs	Avg
Max	93.3	86.6	77.7	72.2	82.5
Mean	94.4	90	77.6	78.8	85.3
Skew	92.2	73.3	74.4	78.8	79.7
Kurt	92.2	73.3	75.5	83.3	81.1
StDev	**98.8**	**98.8**	86.6	88.8	93.3
Avg	94.72	87.22	79.16	79.72	

Taula 2. *Recognition Rate* (%) per al *dataset* Weizmann per a $\beta = 2$.

	$\beta = 4$				
Method	PHOW	LIOP	HOG	SMFs	Avg
Max	93.3	88.8	76.6	74.4	83.3
Mean	96.6	90	80	76.6	85.8
Skew	94.4	74.4	72.2	76.6	79.4
Kurt	93.3	74.4	72.4	74.7	78.7
StDev	**98.8**	**98.8**	87.7	84.4	92.5
Avg	95.8	88	79.2	78	

Taula 3. *Recognition Rate* (%) per al *dataset* Weizmann per a $\beta = 4$.

	β = 6				
Method	PHOW	LIOP	HOG	SMFs	Avg
Max	94.4	86.6	74.4	73.3	82.2
Mean	95.5	88.8	78.8	74.4	84.4
Skew	93.3	75.5	72.1	77.7	79.7
Kurt	92.8	75	71.6	77.2	79.1
StDev	**98.8**	96.6	88.8	84	92.1
Avg	95.5	86.9	78.6	77.4	

Taula 4. *Recognition Rate* (%) per al *dataset* Weizmann per a β = 6.

	β = 8				
Method	PHOW	LIOP	HOG	SMFs	Avg
Max	92.2	88.8	74.4	75.5	82.7
Mean	94.4	92.2	78.8	75.5	85.3
Skew	94.4	75.4	73.3	81.1	81.1
Kurt	93.4	76.5	72.1	83.3	81.3
StDev	**98.8**	**98.8**	90	82.2	92.5
Avg	95	88.9	79.2	78.6	

Taula 5. *Recognition Rate* (%) per al *dataset* Weizmann per a β = 8.

	β = 10				
Method	PHOW	LIOP	HOG	SMFs	Avg
Max	91.1	90	70	75.5	81.7
Mean	94.4	90	78.8	76.6	85
Skew	95.5	76.6	74.4	78.8	81.4
Kurt	95.7	76.8	74.4	80	81.7
StDev	**97.7**	**97.7**	87.7	81.1	91.1
Avg	94.7	88.6	77.7	78	

Taula 6. *Recognition Rate* (%) per al *dataset* Weizmann per a β = 10.

Els resultats mostren que el mig obtingut amb la funció de projecció *StDev* supera a la resta de les funcions a la majoria dels casos. Podem veure que utilitzant la desviació estàndard del valor de les fibres, el mínim està per sobre del 91% i el millor resultat és del 94%. A més, aquestes taules mostren que la funció *StDev* i la tècnica PHOW combinades, aconsegueixen un 98,8% per a β =0, 2, 4, 6 i 8.

S'han obtingut els millors resultats amb 900 *visual words* i la tècnica PHOW.

La figura 37, mostra el rendiment mig de cada funció respecte a β. La funció *StDev* rendeix millor que la resta de funcions, com s'ha comentat anteriorment. A més, es pot veure que les projeccions rendeixen millor per a $\beta = 0$ i per tant, l'increment del veïnat dels tractes no milloren el rendiment mig. Per tant, $\beta = 0$ sembla la millor elecció.

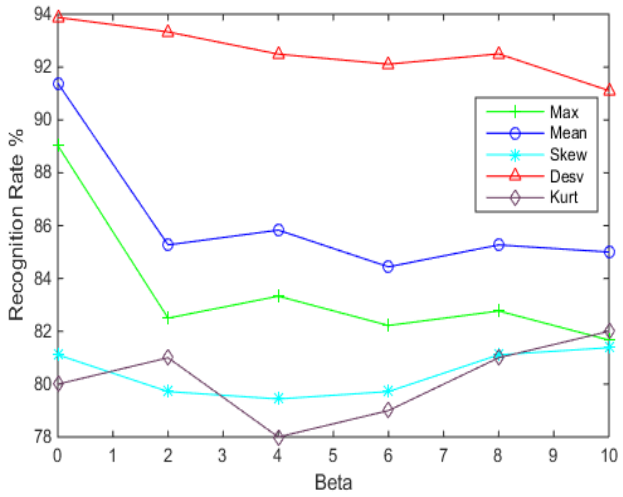


Figura 37. Rendiment mig per a les funcions *Max*, *Mean*, *Skew*, *StDev* i *Kurt* respecte a β per al *dataset* Weizmann.

La figura 38 mostra el rendiment mig de cada tècnica d'extracció/descripció de característiques i el rendiment per a la funció *StDev*. La figura 38 (a) mostra que PHOW rendeix millor que la resta de tècniques. La figura 38 (b) mostra que tot i que PHOW i LIOP obtenen bons rendiments per la funció *StDev*, PHOW rendeix molt bé per a tots els valors de β i per tant és més adequat pels nostres triplets.

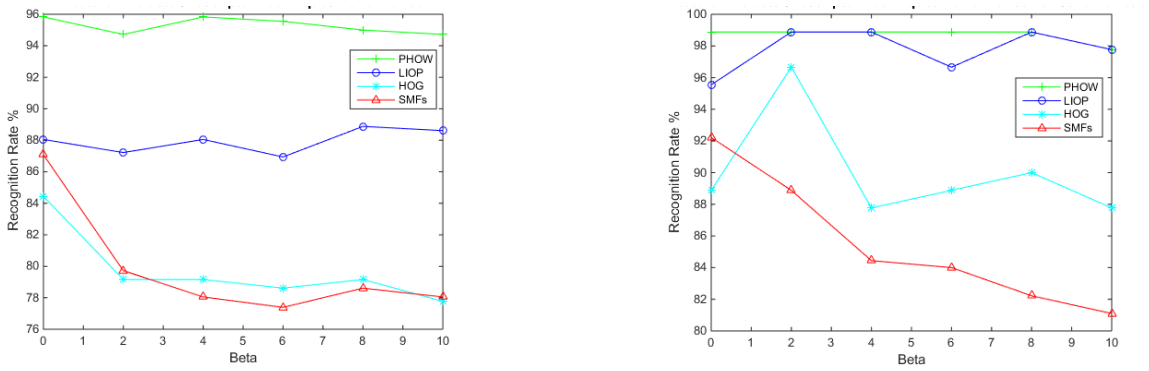


Figura 38. ***a***: Rendiment mig per la funció *StDev* per als extractors/descriptors respecte a β. ***b***: Rendiment mig per la funció *StDev* respecte a β per al *dataset* Weizmann.

Les taules 7, 8, 9, 10, 11 i 12 mostren els resultats obtinguts al *dataset* KTH per a β=0, 2, 4, 6, 8 i 10 generats utilitzant les funcions *Max*, *Mean*, *Skew*, *Kurt* i *StDev*. Les columnes mostren els resultats obtinguts per cada tècnica d'extracció/descripció de característiques aplicada al triplet. Les files mostren els resultats obtinguts pels triplets generats per cada una de les funcions. La darrera fila i columna mostren el rendiment mig de les tècniques d'extracció/descripció de característiques i el rendiment mig de cadascuna de les funcions respectivament.

Els resultats mostren que el rendiment mig de classificació de la funció *StDev* supera una altre vegada la resta de tècniques per a pràcticament qualsevol valor de β.

Es pot veure que per a la funció *StDev* el mínim rendiment mig és lleugerament per sobre del 90% i el millor rendiment mig per aquesta funció és 93%. A més, les taules 8 i 9, mostren que la funció *StDev* i PHOW combinats aconsegueixen un 97.5% d'índex de reconeixement per a β=2 i β=4.

	β = 0					
Method	PHOW	LIOP	HOG	SMFs	Avg	
Max	95.33	88.5	93.5	84.16	90.37	
Mean	95.33	84	90.66	73.33	85.83	
Skew	84.16	80.66	88	86.66	84.87	
Kurt	84.3	80.5	87.9	86.7	84.85	
StDev	**96.83**	92.16	93.83	89.16	93	
Avg	92.91	86.33	91.5	83.33		

Taula 7. *Recognition Rate* (%) per al *dataset* KTH per a β = 0.

	$\beta = 2$				
Method	PHOW	LIOP	HOG	SMFs	Avg
Max	94.83	88.5	90.66	80	88.5
Mean	94.33	82.33	90.5	75.83	85.75
Skew	91.16	83.66	89	87.5	87.83
Kurt	91	83.1	88.8	87.1	87.5
StDev	**97.5**	90.16	93.66	90	92.83
Avg	94.46	86.16	90.96	83.33	

Taula 8. *Recognition Rate* (%) per al *dataset* KTH per a $\beta = 2$.

	$\beta = 4$				
Method	PHOW	LIOP	HOG	SMFs	Avg
Max	96.16	88.16	89.5	76.66	87.62
Mean	94.16	79.16	90.83	72.5	84.16
Skew	94.83	84.16	88.83	81.66	87.37
Kurt	93.3	84	86.8	80.1	86.05
StDev	**97.5**	89.66	94.5	79.16	90.21
Avg	95.66	85.29	90.92	77.5	

Taula 9. *Recognition Rate* (%) per al *dataset* KTH per a $\beta = 4$.

	$\beta = 6$				
Method	PHOW	LIOP	HOG	SMFs	Avg
Max	95.55	89.16	87.83	81.16	88.41
Mean	94	76.66	89.33	75.83	83.96
Skew	94.66	84.16	88.83	92.5	90.04
Kurt	93.6	83.1	90.9	93.6	90.3
StDev	**97**	91.16	94.83	84.16	91.79
Avg	95.29	85.29	90.21	83.41	

Taula 10. *Recognition Rate* (%) per al *dataset* KTH per a $\beta = 6$.

	$\beta = 8$				
Method	PHOW	LIOP	HOG	SMFs	Avg
Max	95	88	97.16	79.16	87.33
Mean	93.33	79	89	73.33	83.67
Skew	94.83	83.83	89.33	90.83	89.71
Kurt	96	88.8	94	84.4	90.8
StDev	**96.83**	89	94.5	85	91.33
Avg	95	84.96	90	82.08	

Taula 11. *Recognition Rate* (%) per al *dataset* KTH per a $\beta = 8$.

	$\beta = 10$				
Method	PHOW	LIOP	HOG	SMFs	Avg
Max	95.16	89.16	86.16	85.83	89.08
Mean	93.16	75.83	88.66	65.83	80.87
Skew	94	84.66	88.83	85	88.12
Kurt	95.26	89.06	87.1	86.3	89.43
StDev	**97.33**	89.8	93.83	83.33	91.07
Avg	94.91	84.86	89.37	80	

Taula 12. *Recognition Rate* (%) per al *dataset* KTH per a $\beta = 10$.

La figura 39 mostra gràficament el rendiment mig de cada funció respecte al valor de β . La funció *StDev* rendeix millor que la resta de funcions. A més, es pot veure que per $\beta > 0$ no millora el rendiment significativament per a les funcions *Max*, *Mean* i *StDev*, però el rendiment de la funció *Skew* en canvi millora significativament per a $\beta > 0$. La figura 39, mostra que la funció *StDev* i $\beta = 0$ obtenen el millor rendiment mig.

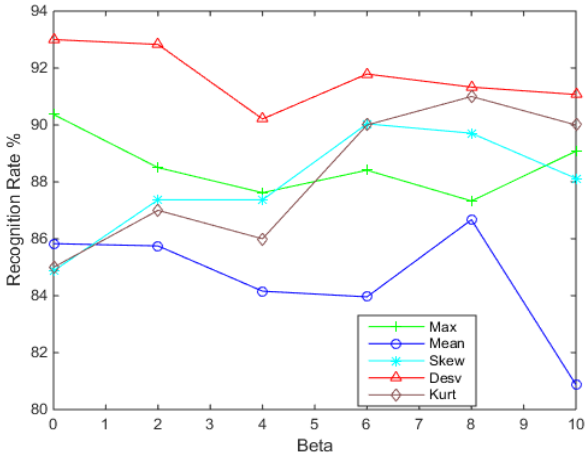


Figura 39. Rendiment mig per les funcions *Max, Mean, Skew, StDev* i *Kurt* respecte a β per el *dataset* KTH.

D'altre banda, la figura 40(a) mostra que PHOW supera la resta de tècniques i per tant aquesta és més adequada pels nostres triplets. La figura 40(b) mostra que PHOW obté el millor rendiment per a la funció *StDev*. S'han obtingut els millors resultats amb 900 *visual words*.

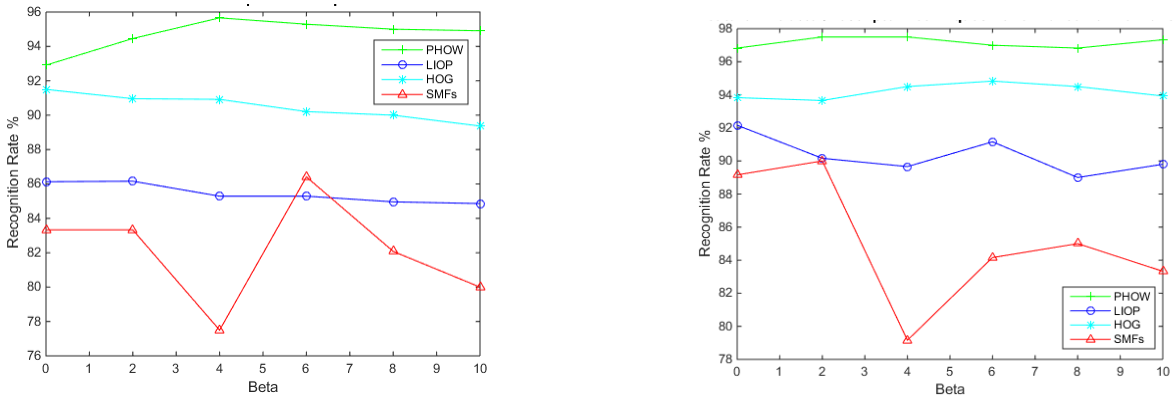


Figura 40. ***a***: Rendiment mig per a la funció *StDev* dels extractors/descriptors respecte a β . ***b***: Rendiment mig per a la funció *StDev* respecte a β per el *dataset* KTH.

4.6.2 Comparativa amb MHI

S'han comparat les nostres projeccions amb una altre mètode de *template* temporal com MHI. De forma similar a l'experiment previ, i amb l'objectiu de comparar les dues tècniques, s'han aplicat les tècniques d'extracció/descripció de característiques HOG, LIOP, SMFs i PHOW sobre les imatge MHI.

Primer, s'ha utilitzat el *dataset* Weizmann per testejar. S'ha fixat β = 0 per aquest experiment ja que aquesta mida de tracte va donar els millors resultats a l'experiment previ. La taula 13 mostra els resultats obtinguts per a cada tècnica d'extracció/descripció de característiques. Es pot veure que el de les nostres projeccions superen a MHI per a totes les funcions testejades.

Method	PHOW	LIOP	HOG	SMFs	Avg
Max	96.66	92.22	82.22	85	89.02
Mean	95.55	91.11	91.11	87.77	91.38
Skew	92.22	73.33	75.55	83.33	81.11
Kurt	91.11	74.44	76.66	84.44	81.66
StDev	98.88	95.55	88.88	92.22	93.88
MHI	88.88	73.33	76.66	81	79.96

Taula 13. Resultats de la comparativa amb MHI en el *dataset* Weizmann per a β = 0.

També s'ha testejat MHI en el *dataset* KTH. De la mateixa forma que la comparativa prèvia, s'ha fixat β = 0 per aquest experiment. La taula 14 mostra els resultats obtinguts per a cada tècnica d'extracció/descripció. Es pot veure que el *recognition rate* mig utilitzant les nostres projeccions supera MHI per a totes les funcions testejades.

Method	PHOW	LIOP	HOG	SMFs	Avg
Max	95.33	88.5	93.5	84.16	90.37
Mean	95.33	84	90.66	73.33	85.83
Skew	84.16	80.66	88	86.66	84.87
Kurt	85.2	80.1	89.4	87	85.42
StDev	96.83	92.16	93.83	89.16	93.0
MHI	90.33	60.83	86	73.33	77.62

Taula 14. Resultats de la comparativa amb MHI en el *dataset* KTH per a β = 0.

4.6.3 Comparativa amb l'estat de l'art per reconeixement d'accions

Per fer la comparativa s'han sintonitzat tots paràmetres dels algorismes als valors reportats pels autors en els seus papers originals utilitzant el *dataset* Weizmann: S'han utilitzat les configuracions de sub-histogrames 2x2x2 i 4x4x4, i histogrames de 8x4 per representar θ i ϕ en el descriptor SIFT 3D [6].

S'ha utilitzat una mida de *codebook* de V = 4000; suport espacial i temporal σ0 = 8, τ0 = 6; número de histogrames cells M = 4, N = 3; número de *supporting mean gradients* S = 3; valor de *cut-off* c = 0.25; i tipus de polièdric icosaedre de orientació completa a [5]. A [14] van utilitzar 500 característiques basades en gradients i van afegir una etapa C3 d'alt nivell a la seva arquitectura jeràrquica. Per a [87], els paràmetres del detector van ser σ = 1.2 i τ = 1.2, la dimensionalitat dels descriptors corresponents van ser 100 i la mida del *codebook* 1200.

Weizmann	**%**	**KTH**	**%**
Scovanner et al. [6]	84.2	PM [15]	97
Klaeser et al. [5]	84.3	TCCA [16]	95
Niebles et al. [87]	90	PT [68]	93.4
Jhuang et al. [14]	98.8	Niebles et al. [87]	83.3
		Ubalde [104]	91.73
Ours	98.8		97.5

Taula 15. *Recognition Rate* (%) en els *datasets* Weizmann i KTH.

Per al *dataset* KTH també s'han sintonitzat tots els paràmetres dels algorismes als reportats en els papers originals dels autors; a [87] els paràmetres del detector van ser σ = 1.2 i τ = 2.5, la dimensionalitat del corresponent descriptor va ser 100 i la mida del *codebook* 1500. [15] va utilitzar un *Product Manifold* i mida dels *frames* de vídeo reduïts a 20x20 píxels. Per a [68] el descriptor *shape-motion* (vector) va ser de 512 dimensions que consisteix de un descriptor de forma de 256 dimensions i un descriptor de moviment de 64 × 4 = 256 dimensions. El valor de *k* a l'algorisme de *clustering k-means* var ser ajustat en un conjunt de validació per *cross-validation* durant l'entrenament, variant *k* des de 200 a 300 i explotant característiques *joint* i *single-shared-mode* per a [16].

La taula 15 mostra els resultats obtinguts al *dataset* Weizmann i KTH. Es pot veure que l'aproximació presentada aquí supera moltes tècniques de l'estat de l'art.

4.6.4 Comparativa amb l'estat de l'art per reconeixement de gestos

En el següent experiment hem provat d'aplicar la nostra tècnica al reconeixement de gestos. Per fer aquest experiment hem comparat els resultats de la nostra tècnica utilitzant PHOW com a extractor/descriptor de característiques i per cadascuna de les funcions *Max*, *Mean*, *Skew*, *StDev* i Kurt amb altres aproximacions en el *dataset* Cambridge Hand-Gesture. A la taula 16 hem llistat els resultats. Els paràmetres utilitzats de

sintonia de cada una de les tècniques són els reportats pels autor als seus papers originals: trajectòries d'entrenament L=3 obtingudes per fusió, igual corba de distància dels punts d'entrenament de cada trajectòria i $\alpha = \beta 1 = \beta 2 = 1$ per [33], explorant característiques *joint* i *single-shared-mode* per a [29], Per a [16] s'han utilitzat correlacions canòniques espacio-temporals dels descriptors SIFT amb transformacions discriminatives (SIFT ST-DCC), per [31] s'ha fet servir la *2D Discrete Cosine Transform* (DCT) conservant les 15 components de freqüències superiors, dividint la imatge en 9 regions rectangulars i solapant 4 píxels entre dos blocs adjacents, per [105] s'ha utilitzat *Tangent bundles* en *manifolds* especials i mida reduïda dels *frames* de vídeo a 20X20 píxels, s'ha utilitzat un *Product Manifold* i mida reduïda dels *frames* de vídeo per [15], i [69] va utilitzar un *Product Manifold* però no dóna detalls de la configuració utilitzada.

La taula 16 mostra que la aproximació presentada en aquest treball per les funcions *StDev* i *Max* rendeixen millor que altres mètodes de l'estat de l'art per a totes les il·luminacions i té una desviació estàndard més baixa.

També es pot veure que el rendiment de les funcions *Max*, *Mean*, *StDev*, *Skew* i Kurt és similar als obtinguts en els experiments anterior de reconeixement d'accions.

Method	**Set1**	**Set2**	**Set3**	**Set4**	**Total**
Yuan et al.[33]	-	-	-	-	82%
Kim et al.[16]	81%	81%	78%	86%	82±3.5%
Kim et al.[29]	-	-	-	-	85±2.8%
Harandi et al.[31]	86%	86%	85%	88%	86.3±1.3%
Lui et al.[105]	88%	84%	85%	87%	86±3%
Lui et al.[15]	89%	86%	89%	87%	88±2.1%
Lui et al. [69]	93%	89%	91%	94%	91.7±2.3%
Ours (*Max*)	95%	95%	97%	96%	95.75±0.95%
Ours (*Mean*)	90%	90%	89%	86%	88.75±1.89%
Ours (*StDev*)	94%	95%	96%	96%	95.25±0.95%
Ours (*Skew*)	80%	76%	70%	71%	74.25±4.64%
Ours (*Kurt*)	78%	78%	69%	73%	75.25±3.77

Taula 16. Comparativa entre la nostra aproximació i l'estat de l'art al *dataset* Cambridge hand-gesture (%)

4.7 Resultats experimentals per datasets complexos

En aquesta secció, es presenten i discuteixen més resultats experimentals de l'avaluació. Tots els experiments han estat testejats en els *datasets* UCF101 i HMDB51. Hem extret les característiques HOG, HOF, MHB, i trajectòries de les seqüències de vídeo utilitzant el codi subministrat pels autors originals i hem creat un GMMs de mida 256. Després hem reduït la dimensionalitat de cada descriptor aplicant PCA amb un factor de 0.5. Les dimensions dels descriptors han estat: 30 per les trajectòries, 96 per HOG, 108 per HOF i 192 per MBH.

4.7.1 Avaluació de representació de característiques

A aquest apartat hem avaluat si la representació Fisher vector (*Fv*) millora els resultats respecte a la representació BOW. Per aquest experiment no hem utilitzat els *datasets* més simples perquè l'alt rendiment assolit en aquests *datasets*, fan difícil percebre la diferència de rendiment entre BOW i *Fv*.

Les taules 17 i 18 mostren els resultats obtinguts per cada projecció I_{xy}, I_{xz} i I_{yz} individualment i combinada, projectant els canals RGB i OF. Hem estudiat el rendiment obtingut per cada projecció individualment i les 3 projeccions combinades sumant les característiques normalitzades amb *power ℓ2-normalized* de cada projecció. La taula 17 mostra els resultats usant la representació *Fv* i la taula 18 els resultats usant la representació BoW. D'aquests resultats podem concloure que la representació *Fv* és més adequada per treballar amb les nostres projeccions, per tant utilitzarem la representació *Fv* per a la resta d'experiments.

	Fv			
function	I_{xy}	I_{xz}	I_{yz}	Combined
S_{mean} *(RGB)*	49.52	42.58	48.24	63.25
S_{StDev} *(RGB)*	50.31	44.32	50.67	65.61
S_{skew} *(RGB)*	26.47	24.85	24.9	38.31
S_{kurt} *(RGB)*	24.58	29.1	27.58	40.12
S_{max} *(RGB)*	47.2	40.15	47.5	62.41
S_{mean} *(OF)*	34.11	53.07	52.85	61.02
S_{StDev} *(OF)*	50.41	52.04	51.88	64.88
S_{skew} *(OF)*	25.36	23.21	22.10	40.52
S_{kurt} *(OF)*	23.57	26.98	27.69	41.25
S_{max} *(OF)*	39.9	55.37	59.97	67.93

Taula 17. Resultats per a I_{xy}, I_{xz} i I_{yz} individualment i combinades en el *dataset* UCF101 per a les vídeo seqüències RGB i OF utilitzant *Fv*.

	BoW			
function	I_{xy}	I_{xz}	I_{yz}	Combined
S_{mean} *(RGB)*	25.24	26.3	27.7	35.21
S_{StDev} *(RGB)*	26.1	27.2	28.9	37.8
S_{skew} *(RGB)*	18.14	17.14	16.4	25.14
S_{kurt} *(RGB)*	17.25	18.54	18.2	23.8
S_{max} *(RGB)*	26	28.67	27.6	34.57
S_{mean} *(OF)*	26.7	27.4	24.1	34.01
S_{StDev} *(OF)*	27.51	28.3	30.1	40.3
S_{skew} *(OF)*	18.9	17.54	19.2	27.02
S_{kurt} *(OF)*	19.1	18.4	16.4	27.23
S_{max} *(OF)*	28.7	28.9	32.4	42.1

Taula 18. Resultats per a I_{xy}, I_{xz} i I_{yz} individualment i combinades en el *dataset* UCF101 per a les vídeo seqüències RGB I OF utilitzant BoW.

4.7.2 Projecció de vista simple *vs* projecció de múltiples vistes

A aquest experiment estem interessats en saber si està justificat calcular $M_{function}$ en comptes de $S_{function}$.

Hem calculat $S_{function}$ i $M_{function}$ en vídeo seqüències RGB i OF utilitzant el *dataset* UCF101. La taula 19 mostra els resultats obtinguts per a cada *split* i el resultat mitjà pel *dataset* UCF101. Per calcular $M_{function,}$ hem generat múltiples *templates* per a cada vídeo, segmentant aquest en múltiples segments sobreposats. Hem utilitzat una finestra temporal de mida 15 i solapament de 5 *frames*.

Method	SPLIT1	SPLIT2	SPLIT3	AVERAGE
S_{mean} (RGB)	63.25	62.45	63.85	63.18
S_{StDev} (RGB)	65.61	66.25	64.85	65.57
S_{max} (RGB)	62.41	61.89	63.58	62.62
S_{mean} (OF)	61.02	60.24	62.35	61.2
S_{StDev} (OF)	64,88	65.2	64.8	64.96
S_{max} (OF)	67.93	65.4	66.87	66.73

Taula 19. Resultats pel mètode $S_{function}$ a les vídeo seqüències RGB i OF en el *dataset* UCF101.

La taula 20 mostra els resultats obtinguts per a cada funció *f* utilitzada. Hem descartat les característiques generades utilitzant la *Kurtosi* i el *Skewness* perquè van obtenir un mal rendiment a l'experiment prèviament realitzat. Podem veure que $M_{function}$ supera a $S_{function}$ per totes les funcions tant a les vídeo seqüències RGB com a les OF.

Method	SPLIT1	SPLIT2	SPLIT3	AVERAGE
M_{mean} (RGB)	69.15	70.54	69.84	69.85
M_{StDev} (RGB)	70.6	73.35	72.35	72.1
M_{max} (RGB)	69.38	66.99	68.53	68.3
M_{mean} (OF)	75.01	72.5	75.69	74.4
M_{StDev} (OF)	69.88	71.28	71.66	70.94
M_{max} (OF)	74.89	75.14	76.56	75.53

Taula 20. Resultats pel mètode $M_{function}$ a les vídeo seqüències RGB i OF en el *dataset* UCF101.

4.7.3 Contribució de les diferents funcions *f*

A aquest experiment estem interessats en conèixer com el rendiment és millorat quan fusionem les *templates* projectades utilitzant diferents funcions *f*. Hem combinat cada funció *f* utilitzant *Sum Pooling*. Per fer això, hem sumat les característiques normalitzades *power ℓ2* de cada funció. La figura 41 mostra el diagrama de blocs de la estructura utilitzada a aquest experiment.

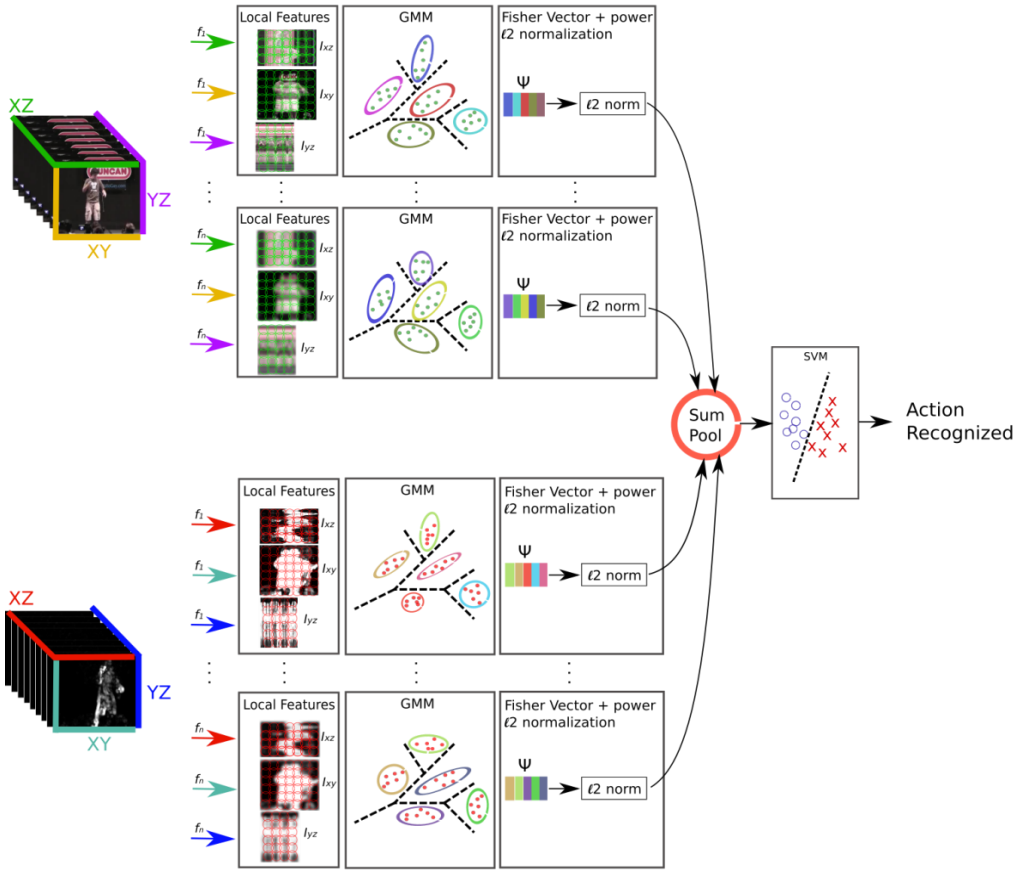


Figura 41. Diagrama de blocs de l'estructura utilitzada a l'experiment.

La taula 21 mostra la millora de rendiment obtinguda quan afegim una de les *templates* calculades amb les diferents funcions *f*. Hem utilitzat *Sequential Forward Selection* (SFS) [66] per afegir les funcions progressivament. D'aquesta forma, hem seleccionat primer la funció que va obtenir el millor resultat individualment (M_{max} (OF)). Per la resta de característiques, hem afegit la característica que obté el millor rendiment quan la combinem amb les característiques prèviament seleccionades. A més a més, com a l'experiment anterior, hem descartat les característiques generades utilitzant les funcions de *Kurtosi* i la de *Skewness* perquè són les que van obtenir el pitjor rendiment quan les hem combinat a les característiques prèviament seleccionades.

Template	UCF101	HMDB51
M_{max} (OF)	75.53	41.1
+ M_{mean} (OF)	76.83	43.26
+ M_{StDev} (OF)	78.14	44.11
+ M_{StDev} (RGB)	79.08	45.2
+ M_{mean} (RGB)	79.81	45.9
+ M_{max} (RGB)	80.1	46.2

Taula 21. Significança de les *templates* en els *datasets* UCF101 i HMDB51.

Podem veure de dalt a baix la millora de rendiment afegida per cada *template*. Tot i que les seqüències OF presenten un rendiment més alt que les RGB, podem veure que les tres funcions RGB sumades obtenen una millora del 2 % aproximadament.

Podem concloure que podem millorar el rendiment combinant diferents funcions f calculades a vídeo seqüències RGB i OF.

4.7.4 Fusió amb IDTs

Fins aquí, hem estudiat com obtenir el millor rendiment del nostre mètode basat en *templates*. Ja que un dels objectius d'aquest treball és millorar el rendiment de IDTs, hem fusionat els nostres descriptors als de IDTs utilitzant la configuració que va obtenir el millor resultat al experiment previ, i hem avaluat la millora de rendiment. Primer, hem comparat el rendiment obtingut per cada descriptor de característiques (HOG, HOF, MHB i els extrets de les *templates*) individualment. Després, hem fusionat tots els descriptors i comparat el seu rendiment amb els dels IDTs originals. Hem fusionat tots els diferents descriptors utilitzant *score level fusion* [36]. Per ambdós tècniques hem utilitzat Fisher *encoding* [36]. Hem extret les característiques HOG, HOF, MHB, i trajectòries de les seqüències de vídeo utilitzant el codi subministrat pels autors originals i hem creat un GMMs de mida 256. Després, hem reduït la dimensionalitat de cada descriptor aplicant PCA amb un factor de 0.5. Les dimensions dels descriptors han estat: 30 per les trajectòries, 96 per HOG, 108 per HOF i 192 per MBH. La figura 42 mostra el diagrama de blocs simplificat de l'estructura utilitzada a aquest experiment.

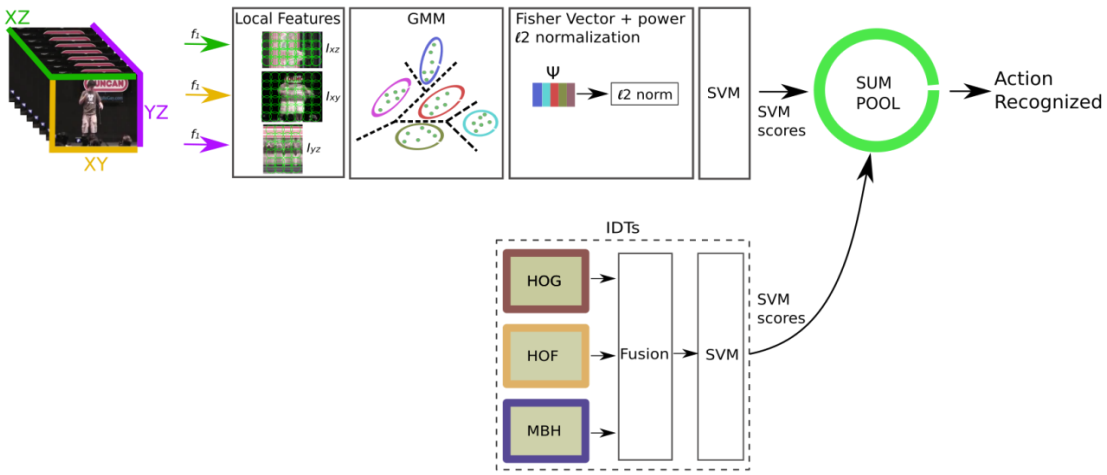


Figura 42. Diagrama de blocs simplificat de l'estructura utilitzada a aquest experiment.

La taula 22 mostra el rendiment obtingut per cada descriptor de característiques individualment i el resultat obtingut quan les combinem. Podem veure que els resultats fent servir *templates* superen les components HOG, HOF i MBH a la majoria dels casos. A més, podem veure que les nostres característiques són complementaries amb les IDTs. Vam obtenir 87% i 60% utilitzant només IDTs als *datasets* UCF101 i HMDB51 respectivament. Per tant, les nostres *templates* porten una millora del 2% i 5% als *datasets* UCF101 i HMDB51 respectivament.

Method	**UCF101**	**HMDB51**
HOG	74,06	45,9
HOF	75,3	45,3
MBHx	77,41	45,25
MBHy	78,12	47,2
$M_{max} + M_{mean} + M_{StDev}$ (OF+RGB)	80,1	46,2
HOG+HOF+MBHx+MBHy (IDTs)	87	60
IDTs+$M_{max} + M_{mean} + M_{StDev}$ (OF+RGB)	89,3	65,3

Taula 22. *Recognition Rate* (%) obtingut amb les característiques IDTs, *templates* i combinant IDTs i els nostres *templates* en els *datasets* UCF101 i HMDB51.

4.7.5 Comparativa amb l'estat de l'art

Hem comparat el nostre mètode amb altres tècniques de l'estat de l'art. Vam mostrar al apartat 4.6.1 que els resultats obtinguts als *datasets* simples com el Weizmann i el KTH van ser molt alts. Per tant,

considerem que calcular IDTs, funcions múltiples i *Fv* no està justificat per aquests *datasets*, ja que els resultats obtinguts utilitzant només descriptors amb una sola funció com *StDev* i representació de característiques com BoW són suficients per superar l'estat de l'art. Els experiments amb aquests *datasets* simples han estat inclosos amb l'objectiu de comparar aquests mètodes que publiquen els seus resultats utilitzant únicament aquests *datasets*.

Per *datasets* més complexos com UCF101 i HMDB51, hem utilitzat la tècnica completa (IDTs+*templates*) per comparar la nostra tècnica amb altres tècniques de l'estat de l'art. La taula 23 mostra els resultats per dos *datasets* complexos com el UCF101 i el HMDB51.

A la comparativa, hem utilitzat la combinació IDTs+ M_{max} + M_{mean} +M_{StDev} (OF+RGB) i la tècnica de representació de característiques *Fv*. La taula 23 mostra les tècniques profundes (CNN) i les tècniques no profundes (no CNN) separadament. Hem considerat els resultats publicats pels autors en els seus papers originals per fer la comparativa.

Els resultats mostren que les nostres 3 projeccions superen moltes de les tècniques de l'estat de l'art quan combinem les *templates* de OF i RGB amb la representació de característiques *Fv*.

Per aquests *datasets*, només [76], [77] i [119] no utilitzen IDTs en els seus resultats finals, la resta d'autors utilitzen o combinen les seves tècniques amb IDTs. [117] combinen a més la seva tècnica amb *Histograms of motion gradients* (HMG) als seus resultats finals. Tot i que les tècniques CNNs obtenen molt bons resultats, creiem que la nostra tècnica no pot ser comparada directament amb elles, ja que aquests mètodes calculen característiques a cada *frame* de RGB i de OF de la vídeo seqüència, mentre que la nostra tècnica només calcula les característiques a les *templates*. La comparativa de la nostra tècnica sense IDTs amb la resta de tècniques, tampoc pot ser comparada de forma justa, ja que les tècniques basades en CNNs utilitzen *streams* exclusivament per l'anàlisi de característiques estàtiques, mentre que les nostres *templates* només consideren informació dinàmica. Pensem que el nostre mètode és només comparable directament amb el de Bilen [78]. Aquest va estudiar diferents escenaris, anomenats Single *Dynamic Image* (SDI) and *Multiple Dynamic Image* (MDI). Aquests són similars als nostres escenaris anomenats $S_{function}$ i $M_{function}$, i per tant, poden ser comparats directament. A SDI, Bilen va experimentar amb les funcions *Max* i *Mean* i les *Dynamic Images* proposades al seu paper. Ells van obtenir 45.4%, 52.6% i 57.9% respectivament, mentre que nosaltres hem obtingut 62.62% i 63.18% en l'escenari $S_{function}$ per les funcions *Max* i *Mean* respectivament. El millor resultat que vam obtenir va ser 65.57% fent servir la funció *StDev* a seqüències RGB i 66.73% fent servir la funció *Max* a seqüències OF. Per MDI, Bilen va obtenir 70.9% fent servir *Dynamic Images* en el *dataset* UCF101, mentre que nosaltres hem obtingut 72.1% en l'escenari $M_{fucntion}$ i la funció *StDev* en imatges RGB i 75.53% fent servir la funció *Max* en vídeos seqüències de OF. Tot i que les tècniques CNN com [76] i [77] obtenen molt bons resultats sense utilitzar IDTs, les *engineered features* com les nostres poden ser encara útils per ser combinades amb *learned features* (CNN) tal com les generades per les CNNs, tal i com van provar a [70]. A més, les

learned features poden obtenir millors resultats que els nostres quan es disposen de moltes mostres d'entrenament, però les *engineered features* encara poden obtenir millor resultats quan només estan disponibles poques mostres a l'entrenament, tal i com va mostrar [108].

UCF101	**%**	**HMDB51**	**%**
		Shallow	
Wu et al. [44]	84.2	Fernando et al.[72]	63.7
Peng et al[36]	87.9	Hoai et al.[73]	60.8
Lan et al.[71]	89.1	Peng et al.[74]	66.8
Peng et al.[43]	87.5	Peng et al.[36]	61.9
		Lan et al.[71]	65.4
Ours	89.3	**Ours**	65.3
		Deep and Very Deep	
Bilen et al.[78]	89.1	Bilen et al.[78]	65.2
Yue-Hei-Ng et al. [76]	88.6	Simonyan et al.[77]	59.4
Simonyan et al.[77]	88		
Bilen et al. [115]	96	Bilen et al. [115]	74.9
Varol et al. [114]	92.7	Varol et al. [114]	67.2
Rohit et al. [116]	93.6	Rohit et al. [116]	69.8
Duta et al. [117]	94.3	Duta et al. [117]	73.1
Feichtenhofer et al. [118]	94.9	Feichtenhofer et al. [118]	72.2
Wang et al. [119]	94.6	Wang et al. [119]	68.9

Taula 23. *Recognition Rate* (%) en els *dataset* UCF101 i HMDB51.

4.7.6 Classificació per projeccions individuals

Als experiments anteriors hem considerat les projeccions I_{xy}, I_{xz} i I_{yz} de forma conjunta per fer la classificació. En aquest experiment fem la hipòtesi de que unes vistes puguin tenir més poder discriminatori que altres, de forma que aquestes tindrien un pes més gran a l'hora de classificar noves mostres. A continuació expliquem el procés per l'entrenament i test.

Procés d'entrenament

Sigui *S* un *dataset* format per *C* classes diferents i *N* mostres de cada classe. Partim el *dataset* en 3 grups. El primer grup seran les dades d'entrenament *t* formades per *N-2* mostres per classe, el segon serà el conjunt de validació *v* format per una mostra per classe i el tercer conjunt serà un conjunt de test *s* format per les mostres restants. Reservarem les mostres *s* "no vistes" per test i utilitzarem les *N-1* mostres sobrants per entrenament i avaluació.

Seleccionem una mostra *v* diferent cada vegada i les *N-2* mostres restants per entrenament.

Per tant, entrenem les SVMs *N-1* vegades per cada mostra de test *s*.

$$t \in \{1..N-2 \mid t \neq v \wedge t \neq s\}.$$

Primer, entrenem cada SVM per calcular el paràmetres *w* i *b* utilitzant *t* mostres i les etiquetes $y_i = 1$ si $t \in C_i$ i $y_i = -1$ per $\forall t \notin C_i$. Repetim aquest procés per a cada classe C_i i per tant obtenim *c* hiperplans.

Una vegada tenim els valors de *w* i *b*, calculem els *scores* de *v*, on cada mostra de *v* pertany a una classe diferent, utilitzant els *c* hiperplans com es mostra a la equació 11.

$$\begin{array}{ccc} \hat{C}_{1,1} = x_1.w_1 + b & . \ . \ . & \hat{C}_{c,1} = x_c.w_1 + b \\ \vdots & & \vdots \\ \hat{C}_{1,c} = x_1.w_c + b & . \ . \ . & \hat{C}_{c,c} = x_c.w_c + b \end{array} \qquad (11)$$

On $\hat{C}_{i,j}$ correspon al *score* obtingut per una mostra corresponent a la classe *i* utilitzant el hiperplà calculat per la classe *j*.

Després, normalitzem els *scores* tal i com mostra l'equació 12.

$$\begin{aligned} \hat{CS}_{ij} &= \hat{C}_{ij} + \left| \min\left(\hat{C}_{ij} \right) \right| \\ \hat{CN}_{ij} &= \frac{\hat{CSj}_{ij}}{\sum_{i=1}^{c} \hat{CS}_{ij}} \end{aligned} \qquad (12)$$

Una vegada normalitzat, tenim $\hat{CS}_{ij} \geq 0$ i $\hat{CN}_{ij} \in \{0..1\}$.

Repetim el procés explicat per *N-1* grups diferents de *t* i una mostra *v* cada vegada. Al final tindrem *(N-1) x C x C* funcions de decisió diferents, és a dir, tindrem $\hat{CN}_{ijk}$ funcions on $1 \le i \le C$, $1 \le j \le C$ i $1 \le k \le N-1$.

Procés de test

De forma similar a la equació 11, però ara utilitzant *s* mostres per test i *N-1* mostres restants per entrenament (abans utilitzades com entrenament i avaluació) amb etiquetes $y_i = 1$ si $t \in C_i$ i $y_i = -1$ si $\forall t \notin C_i$. Repetim aquest procés per cada classe C_i. Una vegada tenim els valors *w* i *b*, calculem els *scores* utilitzant els *c* hiperplans per cada mostra de *s* on cada mostra de *s* pertany a una classe diferent. Posteriorment normalitzem els *scores* aplicant l'equació 12

Per la classificació final comparem dos mètodes lleugerament diferents. El primer es basa en el màxim valor dels productes dels *scores* i el segon en el sumatori del producte dels *scores*.

- ***Mètode 1***

El següent pas correspon a multiplicar cada $\hat{CN}_{ijk}$ obtingut a la fase d'entrenament per cada $\hat{CN}_{ij}$ obtingut a la fase de test i després el màxim valor de cada columna. La equació 13 mostra aquest procés.

$$\hat{E}_{i,j} = \arg\max_{k} \hat{CN}_{ijk} * \hat{CN}_{ij} \qquad (13)$$

On $1 \le i \le C$, $1 \le j \le C$, $1 \le k \le N-1$ i $\hat{E}_{i,j}$ són els índexs corresponents a la classe estimada per cada mostra *i* de cada hiperplà *j*. Finalment, per cada *i* a $\hat{E}_{i,j}$ seleccionem el valor de *j* que ha estat més vegades predit, és a dir, per cada mostra seleccionem la classe més votada per la funció de decisió $\hat{E}_{i,j}$.

Amb aquesta tècnica hem aconseguit un 98.8% de *recognition rate* al *dataset* Weizman utilitzant la funció *Max* i PHOW com a descriptor de característiques. És a dir, hem millorat un 2.2 % respecte a la consideració de les projeccions conjuntes.

- ***Mètode 2***

El següent pas correspon a multiplicar cada $\hat{CN}_{ijk}$ obtinguda a la fase d'entrenament per cada $\hat{CN}_{ij}$ obtingut a la fase de test i després sumar el valors obtinguts per cada producte. L'equació 14 mostra aquest procés.

$$\hat{E}_{i,j} = \sum_{k}^{N-1} \hat{CN}_{ijk} * \hat{CN}_{ij} \qquad (14)$$

Finalment, seleccionem els índexs corresponents als màxims valors de cada columna tal i com mostra l'equació 15.

$$\hat{E}_i = \arg\max_j \hat{E}_{i,j} \qquad (15)$$

On $\hat{E}_i$ és l'índex de la classe predita.

Amb aquesta tècnica hem aconseguit també un 98.8% de *recognition rate* al *dataset* Weizman utilitzant la funció *Max* i PHOW com a descriptor de característiques. És a dir, hem millorat només un 2.2 % respecte a la consideració de les projeccions conjuntes. Per tant, sembla que no està justificat la complexitat afegida per la millora que aporta.

4.8 Conclusions

Hem presentat unes *templates* temporals basades en vistes per reconeixement d'accions. Considerem una vídeo seqüència com un tensor de tercer ordre i calculem tres vistes simples a partir de 3 projeccions diferents de les fibres del subtensor.

A diferència de MHI, les nostres *templates* són calculades partint de les components *Fx* i *Fy* del OF o dels RGB. Quan aquestes són calculades partint de les components *Fx* i *Fy*, poden retenir informació de direcció i magnitud. A més, ja que els triplets han estat calculats a partir de 3 projeccions diferents, aquests poden retenir informació de moviments previs, reduint el problema de l'auto-oclusió. A les projeccions obtingudes, poden ser utilitzades tècniques estàndard d'extracció/descripció de característiques per reconeixement d'accions.

Hem dissenyat alguns experiments amb l'objectiu de testejar quina funció de projecció i quin extractor/descriptor de característiques és més adequat per la nostra aproximació. Hem separat els experiments en dos grups en funció de la complexitat dels *datasets* utilitzats. Els resultats experimentals en els *datasets* més simples han mostrat que utilitzant la desviació estàndard com a funció de projecció, combinada amb PHOW com a descriptor de característiques, dóna el millor índex de reconeixement. A més, els resultats mostren que les projeccions obtingudes rendeixen millor considerant fibres individuals que considerant tractes amb un veïnat.

Les *templates* temporals obtingudes rendeixen millor que les *templates* clàssiques de MHI.

Els resultats experimentals mostren també que les nostres *templates* combinades amb PHOW superen mètodes de l'estat de l'art en reconeixement d'accions.

Tot i que els estudis anteriors han estat fets amb *datasets* de reconeixement d'accions, hem testejat també els resultats amb un *dataset* de reconeixement de gestos i hem comprovat que la tècnica supera a moltes tècniques de l'estat de l'art.

També hem experimentat amb una tècnica de classificació considerant les projeccions individualment. Hem fet l'experiment amb el *dataset* Weizmann i tot i que aquesta tècnica millora el rendiment una 2.2%, creiem que no justifica, per la majoria d'aplicacions, la complexitat i temps de processat que introdueix a la fase d'entrenament i a la fase de reconeixement.

Fent servir dos *datasets* més complexos, hem computat les *templates* a partir de les dues components del OF i també a partir del vídeo RGB fent servir les funcions *Max*, *Standard Deviation*, *Mean*, *Skewness* i *Kurtosi*. Hem utilitzat el descriptor de característiques PHOW, ja que aquest va obtenir el millor rendiment en els experiments amb *datasets* més simples.

Hem experimentat també amb dos diferents escenaris. $M_{function}$ computa múltiples *templates* per cada vídeo seqüència partint aquesta en múltiples segments sobreposats mentre que $S_{function}$ computa una única imatge per cada projecció. $M_{function}$ va obtenir millor rendiment que $S_{function}$.

Els resultats dels experiments han mostrat també que la codificació *Fv* dóna uns resultats més precisos que la codificació *Bag Of Features* (BoF).

Les característiques obtingudes utilitzant diferents funcions de projecció com *Max*, *Mean* i *Standard Deviation* combinades mitjançant *Sum Pooling*, són complementaries i el rendiment augmenta quan són afegides més funcions.

Els descriptors obtinguts de les nostres *templates* obtenen un molt bon rendiment quan són comparats amb HOG, HOF or MHB individualment.

Finalment, hem afegit les característiques de les nostres *templates* a les de HOG, HOF i MBH per millorar el rendiment de la tècnica IDTs.

Els experiments han demostrat que amb *datasets* simples com Weizmann i KTH les *templates* són suficient per superar les tècniques de l'estat de l'art sense necessitat d'utilitzar IDTs. Per *datasets* complexes com UCF101 i HMDB51, fent servir les característiques com a un complement de IDTs, aquests superen moltes tècniques de l'estat de l'art, millorant el rendiment de IDTs en aquests *datasets* en un 2% i un 5% respectivament. Tot i que les darreres tècniques basades en CNNs obtenen molt bon rendiment, aquests no poden ser comparats directament amb la nostra tècnica, ja que aquests calculen característiques a cada *frame* RGB i OF de la vídeo seqüència. D'altra banda, la comparativa de les tècniques sense utilitzar IDTs, tampoc es pot fer de forma justa, ja que les nostres *templates* només consideren informació dinàmica, mentre que la resta de tècniques consideren *streams* exclusivament per la informació estàtica.

5 Discussió

Els resultats dels experiments realitzats amb la tècnica basada en la projecció de subtensors obté un rendiment molt alt. Tot i que aquesta tècnica fa una reducció molt gran de les característiques generades respecte a la majoria de tècniques de l'estat de l'art, un inconvenient d'aquesta és que en el millor dels casos projecta un mínim de tres *templates*, i per tant la capacitat de càlcul necessitada augmenta respecte a tècniques que només projecten un sol template con la plantejada a l'annex A d'aquesta tesi. Tot i així, la necessitat d'un hardware potent, serà molt menor que el necessitat per tècniques que computen característiques a cada *frame* de la vídeo seqüència.

Respecte a les tècniques de *deep learning*, la principal desavantatge de la nostra tècnica és que les basades en *deep learning* poden obtenir un rendiment superior quan disposen de moltes mostres d'entrenament. La majoria de *datasets* de reconeixement d'accions humanes utilitzats actualment a l'estat de l'art contenen accions típiques, con poden ser les efectuades a esports, cuina, etc.. Això fa que es pugui disposar d'un número de mostres d'entrenament que pot ser gran en alguns casos, però per altres tipus d'accions con poden ser gestos, o accions concretes d'un camp específic pot resultar més difícils disposar de la quantitat suficient com per entrenar una CNN i obtenir un alt rendiment. En aquests casos la tècnica basada en projecció de subtensors pot ser una bona alternativa a les tècniques CNN per obtenir un alt rendiment.

La fusió de *engineered features* i *learned features* ha mostrat que millora el rendiment final de la majoria de tècniques. És per això que pràcticament tots els autor mostren els seus millors resultats fusionant les seves tècniques CNNs amb una o inclús en alguns casos dues tècniques no CNN. La millora aconseguida en els darrers anys per les tècniques basades en CNNs -deguda principalment a l'ús d'arquitectures cada vegada més profundes- ha fet que alguns *datasets* com el UCF101 quedin a prop de la saturació, de forma que la fusió amb *engineered features* en aquest *dataset* sembla que cada vegada aporta menys millora al rendiment final. Tot i així, quan s'analitza la millora aportada per la fusió en *datasets* en els que encara els resultats estan lluny de la saturació, es veu que la fusió amb *engineered features* encara aporta una millora important. Això fa pensar que les *engineered features* encara son útils des de el punt de vista de la fusió per millorar els resultats en els *datasets* més complexos actuals i els que possiblement estableixin l'estat de l'art en el futur.

A l'annex A d'aquesta tesi hem plantejat una tècnica que projecta un sol template per cada seqüència. Els experiments han demostrat que es pot obtenir un alt rendiment projectant la vídeo seqüència amb funcions simples com *Max, Mean, StDev*. El rendiment d'aquesta tècnica basada en la transformada de Radon és més limitat que el comentat anteriorment, ja que aquest genera una única *template* per a tota una seqüència d'imatges. És evident que el rendiment d'aquesta tècnica no és comprable amb l'obtingut per les tècniques actuals de *deep learning*, però aquest pot ser útil per entorns on el poder computacional és limitat i per tant es necessita processar el mínim d'informació possible de cada seqüència de vídeo. S'ha

de tenir present que la forma de presentació d'aquesta tècnica a aquesta tesi ha estat en la seva forma més bàsica (BoW i aplicació d'una única funció de projecció cada vegada), i per tant que els resultats son millorables.

6 Conclusions generals

A aquesta Tesi hem presentat unes *templates* temporals basades en vistes pel reconeixement d'accions humanes a seqüències de vídeo. L'objectiu dels *templates* és reduir una seqüència de vídeo a un grup reduït d'imatges que retinguin al màxim la seva capacitat discriminadora, reduint d'aquesta forma el cost computacional necessari per processar cada *frame* de la seqüència.

Les *templates* presentades consideren una vídeo seqüència com un tensor de tercer ordre i calculen tres vistes simples a partir de tres projeccions diferents de les fibres del subtensor aplicant funcions bàsiques com *Max*, *Mean* o *Standard Deviation*. Les projeccions obtingudes, poden ser utilitzades per tècniques estàndard d'extracció/descripció de característiques pel reconeixement d'accions humanes.

Els experiments realitzats han demostrat que les *templates* obtenen un rendiment igual o superior a moltes tècniques de l'estat de l'art que processen tots els *frames* d'una seqüència. A més, la complementarietat demostrada als experiment entre les *templates* generades a partir de les diferents funcions simples, fa que el sistema pugui anar millorant el seu rendiment afegint noves funcions. L'ús de noves funcions més complexes o de noves funcions complementaries és una línia de recerca oberta.

Una altra conseqüència important d'aquests *templates* que ha quedat demostrada als experiments, és que les característiques generades són complementaries a les d'una tècnica actual i exitosa com és IDTs. Aquesta complementarietat fa que la fusió de les característiques d'ambdues tècniques millori el rendiment global, superant a moltes tècniques de l'estat de l'art.

Molt recentment, les tècniques de *deep learning* com les CNNs han obtingut molts bons resultats en diferents àmbits de la Visió per Computador. A l'àmbit del reconeixement d'accions, la comparativa de les tècniques CNN amb la nostra tècnica no es pot fer directament, ja que aquestes calculen característiques a cada *frame* RGB i OF de les vídeo seqüències, mentre que la nostra tècnica només calcula les característiques a les *templates*. A més, les *templates* només consideren informació dinàmica, mentre que les tècniques CNN utilitzen *streams* exclusivament per tractar la informació estàtica.

D'altra banda, les desavantatges de les tècniques CNN són conegudes; necessitat de gran quantitat de mostres d'entrenament, necessitat d'un *Hardware* molt potent, etc. Per tant, les tècniques clàssiques (No CNNs) de Visió per Computador com les presentades a aquesta tesi, són encara útils quan no es disposa d'un gran número de mostres d'entrenament. A més, la majoria d'autor que han utilitzat CNNs en el reconeixement d'accions humanes, han combinat les característiques CNNs amb tècniques com IDTs per millorar el seus resultats finals aprofitant la complementarietat d'aquestes. És també una línia oberta de recerca estudiar la complementarietat de les característiques generades per les *templates* amb les característiques generades per tècniques com les CNNs.

Com a idees derivades de les *templates* plantejades a aquesta tesi, a l'annex A hem presentat també un altre tipus de *templates* generades a partir de la transformada de Radon. Amb aquesta tècnica el rendiment és més baix que amb les *templates* presentades anteriorment, però té l'avantatge de generar menys *templates* per cada seqüència d'imatges. A aquesta tesi només hem explorat la forma més bàsica d'aquesta

tècnica (*Bag Of Words* i una única funció de projecció), però el seu rendiment podria ser millorat fent servir *Fv* i més d'una funció de projecció. A més, les *templates* generades faciliten la segmentació temporal d'accions humanes tal i com hem demostrat als experiments.

7 Treball futur

Als darrers anys, la majoria d'autors han combinat les seves tècniques CNNs amb les característiques IDTs per millorar els resultats. Tal i com va suggerir a [70], les *engineered features* són encara útils perquè poden millorar el rendiment quan són combinades amb les característiques CNNs. En el futur testejarem si les nostres *templates* són complementaries amb les característiques CNNs per millorar els resultats.

També estem actualment treballant en la millora del rendiment del nostre mètode fent servir OF diferencial, múltiples escales a PHOW o unes altres mètodes de codificació com [71] i [74].

D'altra banda, l'increment del rendiment per mitjà de l'afegiment de noves funcions de projecció complementaries es manté com un problema de recerca obert.

Annex A

A. Treballs addicionals derivats d'aquesta tesi

A.1 Introducció

Als capítols anteriors, hem presentat una tècnica que projecta una seqüència de vídeo a tres *templates*. A aquest capítol presentem una aportació addicional d'aquesta tesi consistent en una tècnica que redueix una seqüència d'imatges a una única *template*. Aquesta reducció es basa en una transformada *Rf* plantejada també com una contribució a aquesta tesi i que està basada en la transformada de Radon. Tot i que la seva eficiència no és tan elevada com la presentada als capítols anteriors, sí que facilita l'aplicació a la segmentació temporal d'accions humanes, tal i com es mostra a la darrera part d'aquest capítol.

A.2 Transformada *R*

La transformada *R* està basada en la transformada de Radon [75], aquesta calcula la suma dels valors de la transformada de Radon al quadrat per a totes les línies d'un mateix angle θ. La transformada *R* [27] es pot representar de la següent forma:

$$R(\theta) = \sum_{\rho} g^2(\rho, \theta) \qquad (16)$$

Per tant, el resultat de la transformada *R* és un senyal 1D, on el valor de cada angle representa la suma normalitzada de la intensitat dels píxels per aquest angle. Els avantatges de la transformada *R* és que aquesta proporciona un senyal 1D partint d'una matriu 2D. Aquesta transformada és invariant a la translació, i a més, si es normalitza, també és invariant a escala.

A.3 Transformada R_f

Aquí plantegem una variant de la transformada *R* que pot funcionar millor en determinats casos que la transformada *R* i que a més, permet adaptar el comportament al tipus de problema a resoldre. A continuació mostren la seva forma general:

$$Rf(\theta) = f(g(\rho, \theta)) \qquad (17)$$

On $g(\rho,\theta)$ és la transformada de Radon i f és una funció que podrem triar en funció de cada tipus de problema i que definirà les propietats de la transformada.

A continuació mostrem algun exemples de tipus de funcions *f* que podrien ser utilitzades.

A.3.1 Transformada R_{max}

Aquesta transformada substitueix la potència quadrada de la transformada *R* per una funció *Max* que selecciona el màxim valor absolut (però conservant el signe) per cada angle.

$$R_{max}(\theta) = \begin{cases} \max_\rho(g(\rho,\theta)) & if\ \text{abs}(\max_\rho(g(\rho,\theta))) \geq \text{abs}(\min_\rho(g(\rho,\theta))) \\ \min_\rho(g(\rho,\theta)) & if\ \text{abs}(\max_\rho(g(\rho,\theta))) < \text{abs}(\min_\rho(g(\rho,\theta))) \end{cases} \quad (18)$$

Aquesta transformada és invariant a translació i si es normalitza dividint pel màxim valor de la imatge resultant de la transformada de Radon també és invariant a escalat .

A.3.2 Transformada R_{dev}

Aquesta transformada substitueix la potència quadrada de la transformada *R* per una funció *dev* que calcula la desviació estàndard per a cada angle.

$$R_{dev}(\theta) = dev_\rho(g(\rho,\theta)) \quad (19)$$

Aquesta transformada és invariant a translació.

A.3.3 Transformada R_{mean}

Aquesta transformada substitueix la potència quadrada de la transformada *R* per una funció *Mean* que calcula la mitjana per a cada angle.

$$R_{mean}(\theta) = mean_\rho(g(\rho,\theta)) \quad (20)$$

Encara que aquesta transformada pugui semblar molt similar a la transformada *R* original, aquesta pot interessar quan la funció $g(\rho,\theta)$ tingui valors negatius que interessi ser considerats, ja que la transformada *R* els ignora.

A.3.4 Representació de les transformades R_f

Les propietats de cadascuna de les transformades R_f està en funció de la funció f triada. A part de les diferències respecte a la invariància, també podem trobar diferències en el comportament quan les apliquem a superfícies que poden contenir valor positius i negatius com poden ser superfícies derivades de les components *Fx* i *Fy* del OF que utilitzarem a aquest treball. La figura 43 mostra una comparativa entre les quatre transformades R, R_{max}, R_{dev} i R_{mean} que es consideraran en aquest treball.

Es pot observar que les transformades R i R_{dev} no distingeixen entre imatges positives i negatives, mentre que les transformades R_{max} i R_{mean} si que detecten les diferencies. Aquest fet ens indica que les transformades R_{max} i R_{mean} poden detectar diferents direccions en imatges de OF, mentre que aquesta informació es perd a les transformades R i R_{dev}.

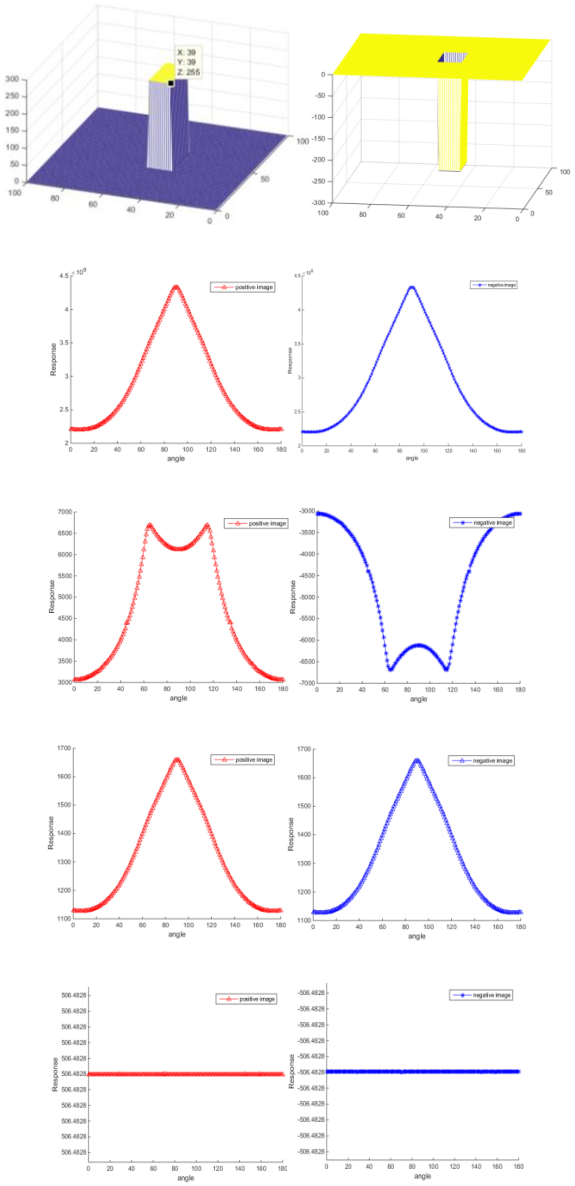


Figura 43. De dalt a baix la resposta de les transformades R, R_{max}, R_{dev} i R_{mean} respectivament a una imatge positiva (esquerra) i una negativa (dreta).

A.4 Aplicació de la transformada R_f al reconeixement d'accions humanes

A aquest capítol s'explica com s'aplica la transformada R_f al reconeixement d'accions humanes en seqüències de vídeo on les accions humanes han estat prèviament segmentades. És a dir, cada acció humana està a una seqüència de vídeo diferent i el sistema només rebrà com a entrada una acció completa cada vegada de principi a fi.

A.4.1 Extracció/descripció de característiques

La transformada *R* ha estat utilitzada al reconeixement d'accions per múltiples autors [25, 90, 91].

La majoria d'aquests autors apliquen aquesta transformada a la imatge de siluetes o a la forma humana extreta de les imatges de la seqüència d'entrada. Això obliga a extreure prèviament la forma o silueta de la persona o persones que realitzen les accions. Els algorismes d'extracció de contorns o forma, acostumen a ser molt sensibles a les condicions d'il·luminació, soroll, etc., de forma que aquesta etapa condiciona de forma important les fases següents.

En aquest treball ens plantegem aplicar la transformada *R* o una de les variants presentades en aquest treball a una seqüència d'accions humanes d'entrada, però en comptes de aplicar-la a les imatges de contorns o forma, l'aplicarem a les imatges del OF calculat prèviament. D'aquesta forma, primer calcularem el OF de cada *frame* d'entrada i posteriorment aplicarem la transformada sobre cada *frame* de OF. Així evitem el càlcul dels contorns o forma i el substituïm pel càlcul del OF que és més robust a canvis d'il·luminació, soroll, etc.

Si apliquem la transformada R_f o una de les seves variants a cada OF *frame* de la seqüència d'entrada, una vegada processada tota la seqüència, tindrem una superfície 2D on cada columna representa la transformada R_f de cada *frame* i cada fila representa un angle θ diferent.

Si apliquem la transformada R_f a cada component *Fx* i *Fy* del OF per separat, tindrem dues superfícies R_x i R_y, a on cada superfície contindrà valors positius i negatius corresponents al moviment amunt i avall o d'esquerra i dreta. La figura 44 mostra un exemple d'aquestes superfícies per a l'acció *bend* del *dataset* Weizmann de cada una de les dues components calculades amb la transformada R_{max}.

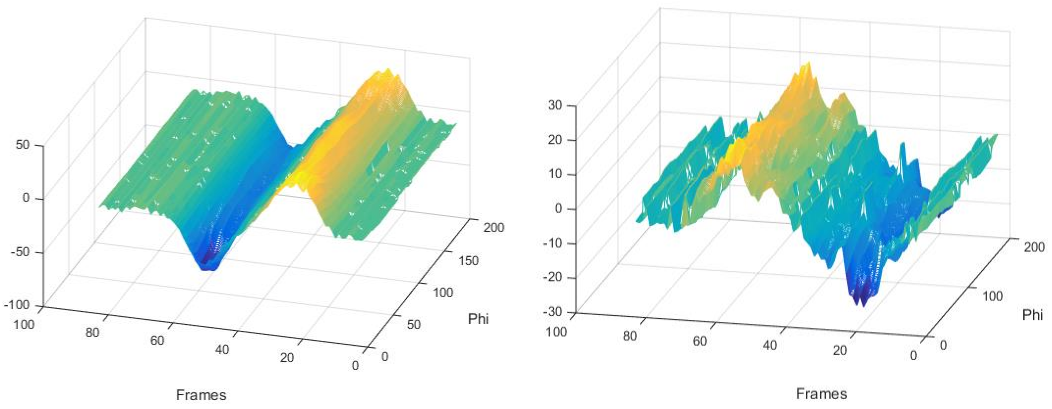


Figura 44. Dos superfícies calculades aplicant la transformada R_{max} a la seqüència de OF de l'acció *bend* del *dataset* Weizmann.

Sobre aquestes superfícies aplicarem un extractor/descriptor de característiques estàndard. En aquest treball, nosaltres hem aplicat PHOW. Aquest mètode deriva del SIFT però es computa de forma densa sobre una graella de *keypoints* definida per la distància entre els *keypoints* en comptes de fer una detecció de *keypoints* prèvia. Ja que aquest mètode no disposa de fase de detecció de *keypoints*, és una bona opció per nosaltres perquè permet aplicar descriptors en zones que poden ser de poc contrast o poc detall, on altres tècniques que sí disposen de detecció de *keypoints* podrien ignorar. La figura 45 mostra un exemple d'aquest descriptor aplicat sobre les superfícies R_f, només es mostren 50 punts de descripció seleccionats aleatòriament.

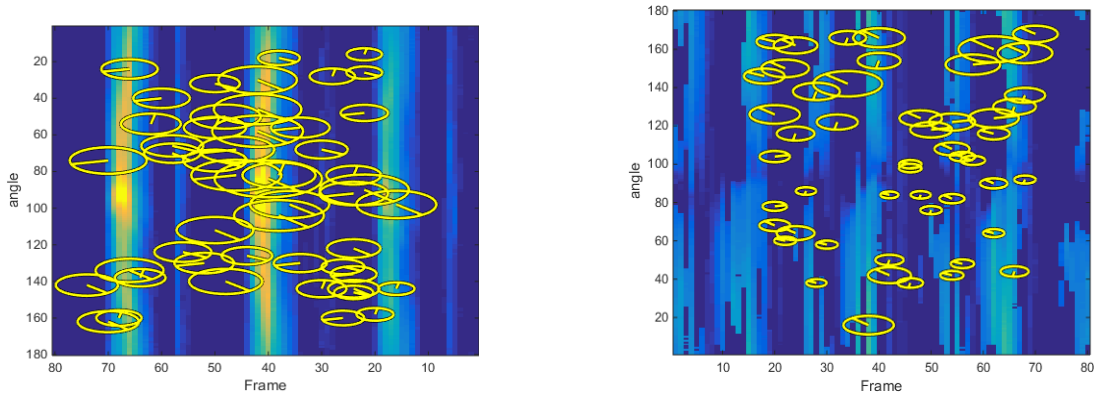


Figura 45. Exemple de PHOW aplicat sobre les superfícies R_f per l'acció *Wave2* del *dataset* Weizmann.

Els cercles estan centrats a la localització dels *keypoints*, els seus radis són les escales, i les línies dintre són les orientacions principals. Es pot veure que les seves orientacions són molt similars entre elles a cada zona de la imatge. Una vegada els *keypoints* han estat localitzats, s'aplica SIFT en la seva forma tradicional en cada *keypoint* , ja que PHOW és simplement un SIFT dens en diferents resolucions.

A.4.2 Classificació d'accions

Una vegada tenim les superfícies resultat de la transformada R_f aplicada a les components *Fx* i *Fy*, utilitzem aquestes com a *templates* espai-temporals per definir cada una de les accions de la mateixa manera que el triplets explicats prèviament. Per poder fer la classificació necessitem una forma de descriure les superfícies que ens permeti comparar les superfícies producte de les accions utilitzades al entrenament amb les testejades en el moment del reconeixement. En aquest treball hem utilitzat la mateixa idea que la utilitzada pels triplets, és a dir, s'ha aplicat un extractor/descriptor de característiques típic del reconeixement d'objectes sobre les superfícies *R*. Ja que les superfícies *R* són molt similars als triplets mostrats en els apartats anteriors, i donat que el extractor/descriptor de característiques que millor

resultat va donar va ser el PHOW combinat amb BoF, aquí també s'ha utilitzat aquest per descriure les superfícies R_f. La figura 46 mostra el diagrama de blocs del mètode complet.

D'aquesta forma, cada seqüència d'entrada serà representada per un conjunt de paraules, on cada paraula descriu una regió de les superfícies R_f.

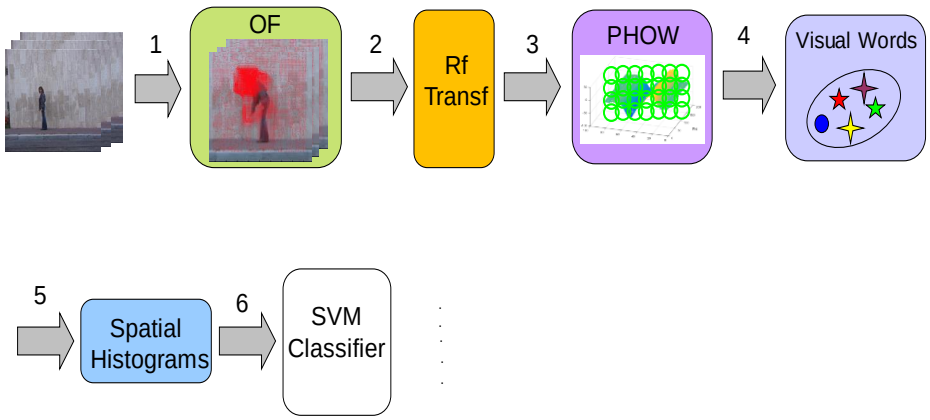


Figura 46. Diagrama de blocs del procés complet utilitzant la transformada *Rf*.

A.4.3 Estudi del problema de l'auto-oclusió

Tot i que les superfícies generades amb la transformada R_f no tenen la capacitat per reduir l'auto-oclusió dels triplets presentat als apartats anteriors d'aquest treball, encara així mantenen una certa capacitat per conservar informació de moviment previ que ha succeït a la mateixa zona de la imatge. A diferència de MHI, les superfícies creades amb les transformades R, R_{max}, R_{dev} o R_{mean} tenen la capacitat de conservar informació de diferents velocitats als moviments que succeeixen a una acció a la mateixa regió de la imatge. La figura 47 mostra un exemple de cadascuna d'aquestes transformades on es pot veure aquest fet. Aquesta figura mostra que en un moviment cíclic con el *wave2* del *dataset* Weizmann, les quatre transformades conserven informació de les diferents velocitat en les direccions *X* i *Y*.

La figura també mostra que les transformades R_{max} i R_{mean} conserven a més informació de direcció del moviment, mentre que les transformades R i R_{dev} perden aquesta informació.

D'altra banda la figura 48 mostra les superfícies generades per cadascuna de les quatre transformades per l'acció *jack* del *dataset* Weizmann. Aquesta acció conté moviment de braços i cames simultàniament i es va servir a l'apartat 3.4 per demostrar com el triplet generat per les tres projeccions podia conservar informació del moviment de les cames i braços i a més la seva direcció. La figura 48 mostra en canvi que tot i que les transformades R_f també conserven informació de direcció del moviment, es fa molt més

complicat distingir si es conserva informació de cames o braços. En el millor dels cassos es pot conservar en *frames* alternatius la informació de cames o braços, però això dependrà de la funció f escollida a la transformada R_f i de la velocitat del moviment de cada part a la imatge. Aquest avantatge del triplet respecte a la transformada R_f no invalida aquesta última per l'ús en àmbits on les accions no requereixen moviments tan complexos o on l'ús de tres *template* pot ser més problemàtic.

La figura també mostra que les transformades R_{max} i R_{mean} conserven a més informació de direcció del moviment, mentre que les transformades R i R_{dev} perden aquesta informació.

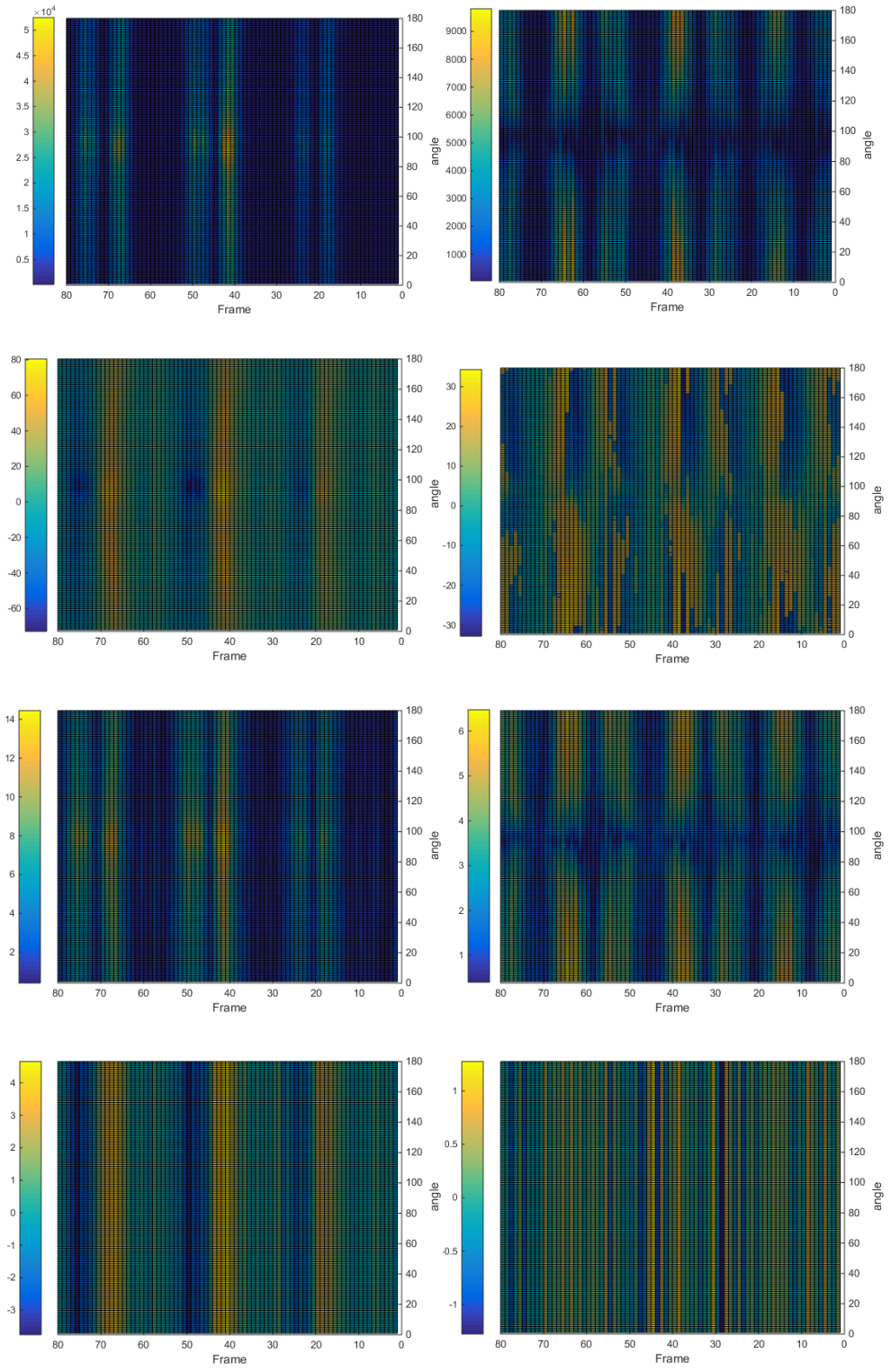


Figura 47. Mostra de les transformades *R*, R_{max}, R_{dev} i R_{mean} aplicades a l'acció *wave2* del *dataset* Weizmann. Esquerra: component *Fx*. Dreta: component *Fy*.

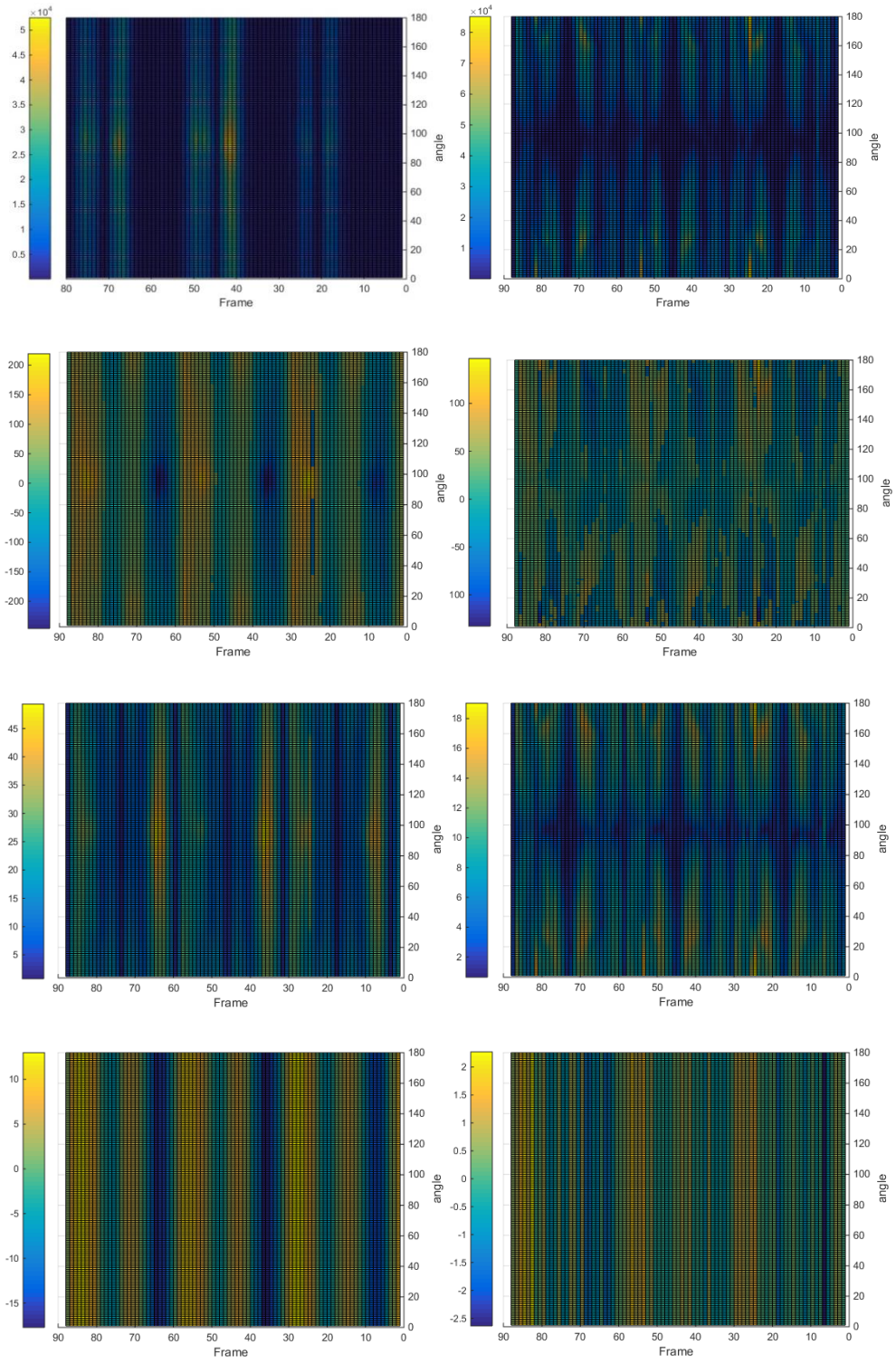


Figura 48. Mostra de les transformades *R*, R_{max}, R_{dev} i R_{mean} aplicades a l'acció *jack* del *dataset* Weizmann. Esquerra: component *Fx*. Dreta: component *Fy*.

A.4.4 Experiments

A.4.4.1 Objectiu dels experiments

L'objectiu del primer experiment és comparar el rendiment obtingut per la transformada *R* i les tres variants de la transformada R_f (R_{max}, R_{mean}, R_{StDev}) per al nostre problema de reconeixement d'accions, on les accions a diferents seqüències de vídeo ja han estat segmentades prèviament, és a dir, en la forma original del *dataset* Weizmann. El segon experiment consisteix en comparar el rendiment de la nostra tècnica amb altres tècniques de l'estat de l'art. Pels dos experiments hem utilitzat el *dataset* Weizmann .

Com a extractor/descriptor de característiques hem utilitzat PHOW. La implementació completa utilitzada és la que reflexa la figura 46.

A.4.4.2 Criteris d'avaluació

Per a cada transformada (R, R_{max}, R_{mean}, R_{StDev}) es calcula el *recognition rate* $\overline{RR}$ de la següent forma:

$$\overline{RR} = \frac{\#samples \quad correctly \quad classified}{Total \quad samples \quad Tested} \qquad (21)$$

Una vegada s'ha determinat la transformada que millor rendeix, aquesta ha estat utilitzada al segon experiment. Al segon experiment, s'ha aplicat novament el mateix extractor/descriptor de *keypoints* (PHOW), i comparat el seu rendiment contra altres tècniques de l'estat de l'art. Per fer la comparativa, s'han considerat els resultats publicats pels autors als seus papers originals.

A.4.5 Resultats experimentals

A la nostra implementació hem utilitzat la tècnica de descripció de característiques PHOW sintonitzada de la següent forma. Per a PHOW s'ha utilitzat *step* = 1, *size*=3, on *step* és la distància entre els *keypoints* i *size* és la mida del descriptor. S'ha testejat *codewords* des de 100 a 1100 paraules visuals per a PHOW. Per a les tècniques d'altres autors, s'han sintonitzat tots els paràmetres dels algorismes als valors reportats

pels autors en els seus papers originals utilitzant el *dataset* Weizmann: S'han utilitzat les configuracions de sub-histogrames 2x2x2 i 4x4x4, i histogrames de 8x4 per representar θ i ϕ en el descriptor SIFT 3D [6].

S'ha utilitzat una mida de *codebook* de V = 4000; suport espacial i temporal σ0 = 8, τ0 = 6; número de *cells* histogrames M = 4, N = 3; número de *supporting mean* gradients S = 3; valor de *cut-off* c = 0.25; i tipus de icosaedre polièdric de orientació completa [5]. A [14] van utilitzar 500 característiques basades en gradients i van afegir una etapa C3 d'alt nivell a la seva arquitectura jeràrquica. Per a [87], els paràmetres del detector van ser σ = 1.2 i τ = 1.2, la dimensionalitat dels descriptors corresponents van ser 100 i la mida del *codebook* 1200.

A.4.5.1 Comparativa de funcions R_f

La taula 24 mostra els resultats obtinguts per les diferents transformades. Es pot veure que la transformada R_{max} obté el millor resultat. També queda clar que la funció *Mean* obté un rendiment molt per sota de la resta de transformades.

Funció Rf	***%***
R	95,55
Max	**98,88**
Mean	88,55
dev	95,55

Taula 24. Resultats al *dataset* Weizmann per a cadascuna de les funcions R_f.

A.4.5.2 Comparativa amb altres tècniques de reconeixement d'accions

La taula 25 mostra els resultats obtinguts al *dataset* Weizmann. Només comparem els resultats obtinguts amb la transformada R_{max}, ja que és la que va obtenir els millors resultats a l'experiment anterior.

Method	%
Scovanner et al. [6]	84,2
Klaeser et al. [5]	84,3
Niebles et al. [87]	90
Jhuang et al. [14]	98,8
Ours	**98,8**

Taula 25. Resultats al *dataset* Weizmann

Es pot veure que l'aproximació presentada aquí supera moltes tècniques de l'estat de l'art. Tot i que [14] obté el mateix rendiment que la nostra tècnica, s'ha de considerar que aquesta és una tècnica bioinspirada que no pot funcionar en temps real. A la nostra tècnica en canvi, el temps mig de computació per reconèixer una acció a una seqüència de 100 *frames*, on cada *frame* té una mida de 160X120 píxels és 900 ms. Computat a un Intel 3.1GHZ i3. El codi ha estat implementat en Matlab i no ha estat optimitzat.

A.4.6 Conclusions

Hem presentat una *template* temporal basada en vistes per reconeixement d'accions. La nostra *template* es calcula aplicant la transformada R_f o una variant d'aquesta a cada *frame* d'una vídeo seqüència.

Hem presentat una forma general de transformada R_f basada en la transformada de Radon que de forma similar a la transformada *R* quan s'aplica a un *frame* torna un senyal 1D.

Triant la funció *f* de projecció correcta, aquesta es pot adaptar a un problema concret. Hem testejat tres transformades creades projectant la transformada de Radon utilitzant les funcions (*Max*, *Mean* i *StDev*) anomenades R_{max}, R_{mean}, i R_{StDev}. La nostra *template* es calcula a partir de les components *Fx* i *Fy* del OF, i pot conservar informació de direcció i magnitud. Hem aplicat un descriptor estàndard de característiques com PHOW i estudiat quina transformada és la més adequada per la nostra *template*. Els resultats experimentals han demostrat que la transformada R_{max}, supera a la transformada R_{mean}, R_{StDev} i també a la transformada *R* original.

Els resultats experimentals han demostrat també que la nostra *template* generada utilitzant la transformada R_{max} i combinada amb PHOW com a descriptor de característiques, supera moltes tècniques de l'estat de l'art.

A.5 Aplicació de la transformada R_f a la segmentació temporal

A aquest capítol s'explica com s'aplica la transformada R_f al reconeixement d'accions humanes en seqüències de vídeo on les accions humanes no han estat prèviament segmentades. És a dir, cada acció humana està encadenada amb la següent sense separació entre elles. Per tant, el sistema rebrà com a entrada una única seqüència amb moltes accions humanes encadenades.

A.5.1 Introducció

En els casos reals de reconeixement d'accions humanes, s'encadenen un conjunt d'accions de forma continua, per exemple una persona que fa esport podria començar a córrer i ajupir-se per seguidament desplaçar-se de forma lateral, etc.. Això fa que no tinguem constància de quan comença i acaba cada acció, i per tant les tècniques presentades prèviament no són directament aplicables als casos reals de reconeixement d'accions humanes. Per millorar aquesta situació, aquí presentem una millora de la tècnica basada en la transformada R_f que pot funcionar en casos de seqüències d'accions encadenades.

El model de segmentació temporal utilitzat a aquest treball és un model basat en la tècnica de finestra lliscant. L'idea és anar desplaçant una finestra de *n*x*m* píxels per les superfícies R_f i reconèixer a quina classe pertany la regió acotada per la finestra en cada moment.

Donada una superfície R_f, i donada una subsuperficie R_{ij} de R_f que conté l'acció *A*, on *i* és el *frame* inicial i *j* el darrer *frame* de l'acció *A*. Considerant $j = i + t$ on *t* és el número de *frames* de l'acció, calculem un conjunt de descriptors PHOW $D[d_1,.....,d_n]$ en una finestra W_{ij} sobre la superfície R_f.

És evident que la finestra ha de ser més estreta que la seqüència més curta que volem reconèixer, de forma que la finestra sempre conté com a màxim una subseqüència de la seqüència a reconèixer.

Donat això, no podem fer l'entrenament amb les seqüències completes que volem reconèixer, ja que en el reconeixement la finestra mai contindrà la seqüència completa.

Per superar aquest problema, s'ha entrenat el sistema amb subseqüències de la mateixa llargada que es farà servir al reconeixement, és a dir, donat una mida de finestra, aquesta mida es fa servir a l'entrenament en regions seleccionades aleatòriament de cadascuna de les seqüències d'entrenament. Posteriorment al reconeixement es desplaça una finestra de la mateixa mida utilitzada a l'entrenament per la seqüència d'entrada i s'alimenta el classificador amb la informació de cada nova finestra d'entrada. La figura 49 mostra un exemple del desplaçament de la finestra per la superfície R_f.

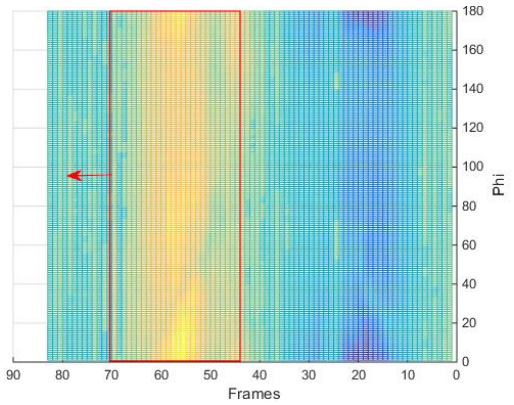


Figura 49. Una finestra de 180x25 píxels lliscant sobre una superfície R_f.

A.5.2 Extracció/descripció de característiques

Primer de tot, projectem la seqüència d'entrada completa a una única *template* utilitzant la transformada R_f. En comptes d'utilitzar la seqüència *raw* de vídeo com a entrada a la transformada R_f, apliquem la transformada R_f a les dues components del OF, obtenint dues superfícies R_{fx} i R_{fy}. Aquestes superfícies poden ser considerades com *templates* espai-temporals definint una seqüència d'acció. Un exemple d'aquesta superfície es mostra a la figura 49. Ja que R_{fx} i R_{fy} han estat calculades aplicant la transformada R_f a cada *frame* de la seqüència de vídeo, cada coordenada en el eix *x* correspon a un *frame* de la seqüència de vídeo, mentre que cada coordenada a l'eix *y* correspon al valor de la direcció ρ de la projecció *Rf* (de 0 a 180º).

El OF ha estat calculat utilitzant l'algorisme de temps real presentat a [92]. Hem concatenat R_{fx} i R_{fy} per obtenir una única superfície per a les etapes de classificació i entrenament. Per tant, obtenim una *template* de *F* x*360* per a una seqüència completa, amb *F* sent el número de *frames* de la seqüència.

La nostra tècnica de segmentació temporal està basada en el model de finestra lliscant. Desplacem una finestra de *N* x *M* píxels sobre les superfícies R_f i realitzem un procés de reconeixement dintre de l'àrea de la finestra, on *N* és l'ample de la finestra (és a dir, el número de *frames* de la seqüència), i *M* és el rang que la finestra abasta (és a dir 360º, el rang complet de direccions). Òbviament, l'ample de la finestra *N* ha de ser més estret que l'acció més curta a ser reconeguda, es a dir $N<F$. Hem triat $N=25$ per les etapes d'entrenament i reconeixement.

Computem un conjunt de *n* descriptors $D[d_1,.....,d_n]$dintre de cada finestra de la superfície R_f. A l'etapa de reconeixement, desplacem una finestra amb la mateixa mida (25x360) sobre la superfície d'entrada per cada *frame* d'entrada. La figura 50 mostra un exemple de finestra lliscant sobre una superfície R_f, i PHOW aplicat dintre d'aquest fragment de la superfície. Hem treballat a una única escala. El número de *keypoints* extrets dintre de la finestra depèn de la mida de la finestra (*N* x*M*), la distància (*D*) triada a la

grid del SIFT dens, i la mida de l'escala (*Sc*) dels descriptors triats. He fet servir *N=25*, *M=360*, *D=1* i *Sc=3*.

Després d'aquest càlcul, hem utilitzat la tècnica BoW. Per fer això, els descriptors similars són agrupats amb l'algorisme *k-means*. Els centres d'aquest *clusters* defineixen un *Visual Codebook*.

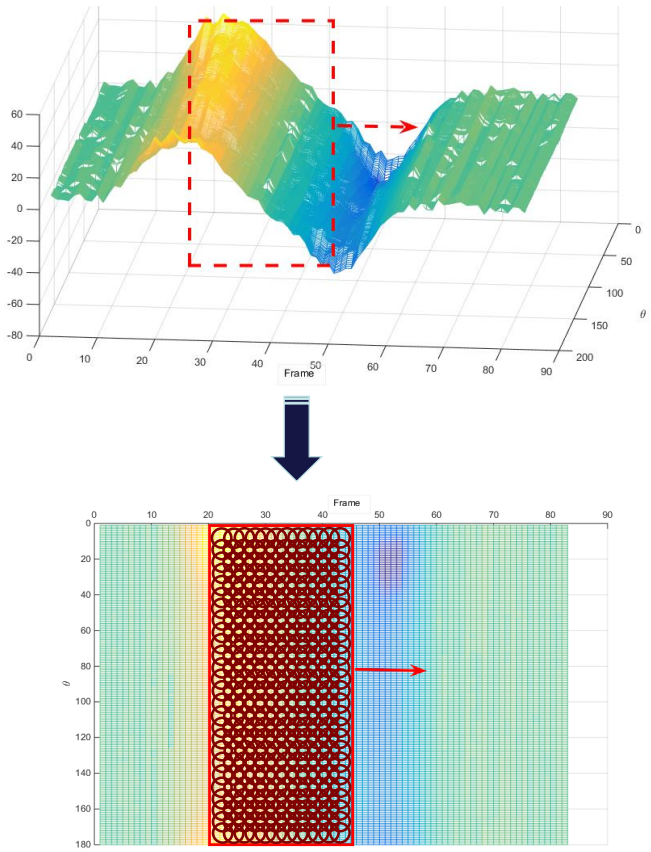


Figura 50. Exemple d'una superfície de projecció calculada fent servir la transformada *Rmax* per l'acció *bend* del *dataset* Weizmann. La seqüència conté 84 *frames*. A dalt: Finestra lliscant sobre una superfície *Rmax*. A baix: PHOW aplicat sobre la mateixa superfície. Els cercles mostren el *keypoints* densament aplicats dintre de la finestra.

A.5.3 Classificació d'accions

La classificació en aquesta tècnica està basada en la resposta dels *scores* de diferents SVMs entrenades per cada una de les classes a reconèixer. L'idea és que una SVM dóna una resposta més alta quan s'alimenta amb un vector de característiques de la mateixa classe per a la que ha estat entrenada. D'aquesta manera, si entrenem una SVM per cadascuna de les classes que pretenem reconèixer, aquesta respondrà de forma més enèrgica quan se li presenti un patró de la classe per a la que ha estat entrenada, mentre que la resta de SVMs tindran una resposta baixa i similar entre elles. De fet, a l'entrenament cada SVM s'entrena amb mostres etiquetades com a positives les que volem que reconegui i com a negatives

les de la resta de classes. D'aquesta manera i de forma similar a [60], s'entrena una SVM lineal per a cada classe utilitzant la tècnica *one-versus-all*, per tant tindrem un conjunt de *Q* classificadors SVM lineals *f1 (Wij), … ,fn(Wij)*, on *Q* és el número de classes.

Donada una acció *A* formada per *N frames*, es seleccionen *P* finestres W_s en localitzacions aleatòries *s* a l'etapa d'entrenament, i s'aplica el extractor/descriptor de característiques PHOW en cada finestra. Es repeteix aquest procés per cada mostra d'entrenament. Així, al final tindrem *C x P* mostres d'entrenament, on *C* és el número de accions d'entrenament i *P* és el número de finestres aleatòries utilitzades a cada acció.

Finalment, s'entrena una SVM per a cada classe d'accions, considerant l'etiqueta 1 si $A_s = A_r$ i -1 a la resta.

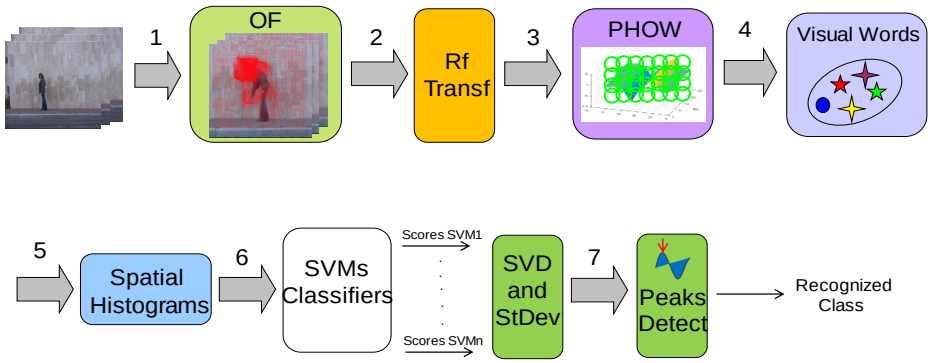


Figura 51. Diagrama de blocs del procés de segmentació temporal complet.

La figura 51 mostra el diagrama de blocs del mètode complet.

Per determinar on comença i acaba una seqüència, utilitzarem la resposta de les SVMs entrenades a la fase d'entrenament. L'interès està en els valors dels *scores* de les SVMs. Així, *scores* alts indiquen proximitat a una classe mentre que *scores* baixos indiquen llunyania a una classe. Per tant, s'espera que els *scores* de la SVM_i serà més gran per a l'acció A_i quan la finestra està al mig de la regió de l'acció i baixa a la resta de SVMs. Esperem que els *scores* de la SVM_i seran pròxims a 1 i a la resta SVM_j serà pròxim a -1 per qualsevol $j \neq i$ quan la finestra està només sobre l'acció A_i.

Quan la finestra està a sobre de regions de no-accions o entre varies accions, la resposta de totes les SVM seran similars.

Ja que les sortides dels *scores* contenen molt de soroll, s'ha aplicat SVD per atenuar el soroll. Per fer això, s'ha aplicat SVD als *scores* de sortida i després s'han eliminat tots els valors singulars menys el

primer. Finalment es fa la inversa del SVD per obtenir el senyal dels *scores* filtrats. La equació següent mostra com la matriu *M* de *n*x*m* elements es pot factoritzar utilitzant SVD:

$$M = U\Sigma V^*$$

On U és una matriu unitària de *m*x*m*, Σ és una matriu diagonal de números reals no negatius de *m*x*n* elements i V^* és una matriu unitària de *n*x*n* elements.

Per segmentar les accions, primer s'ha calculat la desviació estàndard dels *scores* filtrats. Un màxim local indicarà que una acció ha estat detectada i mínims locals indicaran que l'acció ha finalitzat. La figura 51 mostra una exemple dels *scores* filtrats i la desviació estàndard calculada sobre aquest senyal. Per detectar els màxim i mínim locals es fan servir tècniques de detecció de pics basats en la primera derivada. Si es detecta més d'un pic a la mateixa resposta d'una SVM, només es considera una.

A.5.4 Experiments

A.5.4.1 Objectius dels experiments

L'objectiu del primer experiment és comprovar el rendiment de la tècnica aquí presentada en reconèixer accions en una seqüència de vídeo formada per diferents accions encadenades. La sortida de la tècnica utilitzada en aquest experiment serà l'acció reconeguda en cada moment en la seqüència d'entrada.

El segon experiment consisteix també en reconèixer unes accions determinades en una seqüència de vídeo en la que hi ha encadenades diferents accions. A diferencia de l'experiment anterior, en aquest cas la sortida del sistema no serà l'acció reconeguda, sinó si l'acció és una acció coneguda pel sistema o no, però no quina acció concreta és. És a dir, el sistema ens dirà si l'acció és una acció coneguda o és una seqüència de moviments desconeguts pel sistema.

Aquesta implementació estaria pensada per sistemes que una vegada detectada una acció d'interès, apliquessin una altre tècnica per determinar quina acció concreta és.

A.5.4.2 Criteris d'avaluació

Per avaluar els resultats, al primer experiment entrarem una seqüència del *dataset* Weizmann amb les 10 accions encadenades i testejarem que la seqüència de classes reconegudes correspon amb l'ordre d'entrada. Ja que poden donar-se els casos de no detecció d'una acció, error en l'acció reconeguda o

doble detecció d'una acció, comprovarem visualment que les accions reconegudes corresponen al segment de la seqüència d'entrada correcta.

Per al segon experiment el procediment serà el mateix que l'utilitzat anteriorment. En aquest cas la sortida seran dues etiquetes, una d'acció reconeguda o una altre d'acció no reconeguda. Aquí assumim que no tindrem dos seqüències conegudes seguides, de forma que sempre a una acció coneguda seguirà una acció desconeguda, ja que el sistema detectaria tot el fragment com una sola acció reconeguda.

El càlcul del *recognition rate* $\overline{RR}$ pels dos experiments el calcularem de la següent forma:

$$\overline{RR} = \frac{\#\,samples \quad correctly \quad classified}{Total \quad samples \quad Tested} \tag{22}$$

A.5.5 Resultats experimentals

A.5.5.1 Segmentació i reconeixement d'accions

Ja que en aquest experiment l'objectiu és avaluar la capacitat de segmentar accions en una seqüència d'accions concatenades i reconèixer quina es l'acció en el mateix procés, hem entrenat el sistema amb 6 finestres en posicions aleatòries de cada seqüència d'entrenament de 25 píxels d'amplada. A la fase de test s'ha concatenat cada vegada una seqüència amb les 10 accions del *dataset* Weizmann i s'han detectat les accions per la detecció dels màxims i mínims locals a la desviació estàndard. La figura 52 mostra els resultats per una seqüència amb les 10 accions concatenades del *dataset* Weizmann. Quan dos màxims o mínims locals consecutius pertanyen a la mateixa classe, es fusionen en un sol valor.

En aquest experiment hem avaluat dos mètodes de segmentació lleugerament diferents. El primer mètode mesura quina SVM dóna un *score* més alt entre dos mínims locals, la classe corresponent a aquesta SVM serà la que correspon a l'acció reconeguda. Aquest mètode ha obtingut un rendiment del 96.6%

El segon mètode en comptes de mesurar el màxim *score*, el que fa és mesurar el *score* que supera durant més temps a la resta entre dos mínims locals. Aquest mètode ha obtingut un rendiment del 95.5%.

La detecció per màxims locals no ha mostrat bons resultats, així que l'hem descartat en aquest treball.

Tot i que el *dataset* Weizmann és un dels més utilitzats en la literatura de reconeixement d'accions, la majoria d'autors mesuren els seus resultats sobre les seqüències pre-segmentades, de forma que no són

directament comparables. Uns dels pocs autors dels que tenim constància d'haver utilitzat aquest *dataset* sense pre-segmentació són [58], que va reportar un índex de reconeixement del 88.9% i [61] que va reportar un índex de reconeixement del 87.7%.

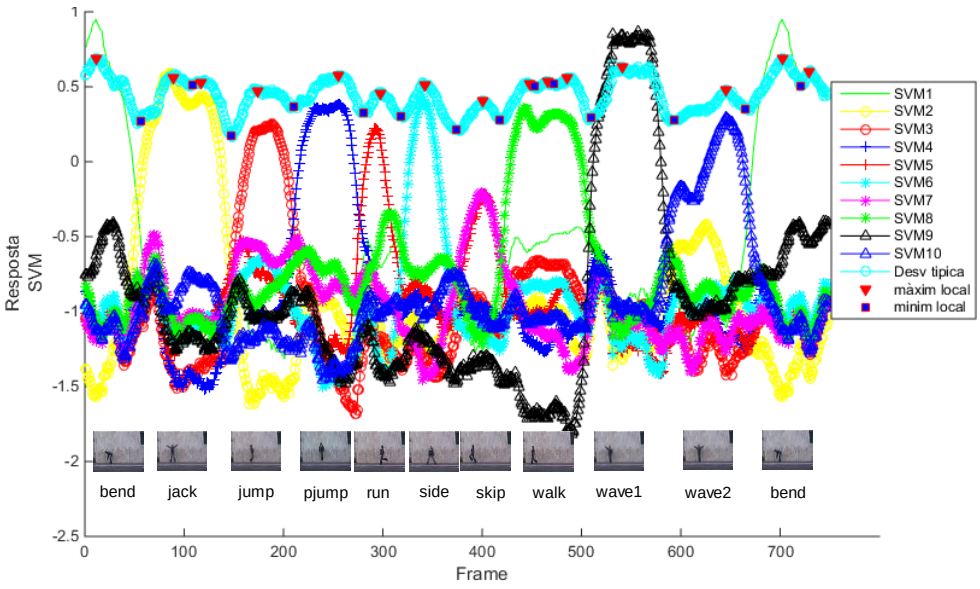


Figura 52. Mostra de la resposta de les SVMs a una seqüència completa del *dataset* Weizmann.

A.5.5.2 Segmentació d'accions respecte altres moviments

L'objectiu d'aquest últim experiment, és avaluar la capacitat de segmentació d'accions d'interès respecte d'altres accions o moviment de persones o objectes no desitjats. Aquest mètode pot ser útil quan disposem d'un mètode de reconeixement d'accions robust, però que no permet la segmentació temporal d'accions, de forma que amb aquest mètode podem determinar en quins *frames* comença i acaba l'acció i posteriorment fer el reconeixement amb un altre mètode.

Per avaluar la tècnica hem considerat com a accions d'interès les accions del *dataset* Weizmann, i com accions de no interès fraccions de les seqüències d'accions o moviments del *dataset* Hollywood que no incloguin les accions realitzades al *dataset* Weizmann. Aquest últim *dataset* conté moltes seqüències d'objectes i persones en moviment molt variades en una mateixa seqüència.

L'entrenament s'ha fet amb el mètode *leave-one-out cross validation*, de forma que com a mostres positives s'han utilitzat totes les mostres del *dataset* Weizmann etiquetades com una sola classe menys les d'una persona que serà utilitzada com a test. En el cas del *dataset* Hollywood, s'han utilitzat seqüències

d'accions a l'entrenament diferents a les utilitzades a la fase de test. Totes les seqüències del *dataset* Hollywood utilitzades al entrenament, s'han etiquetat com una mateixa classe.

La fase de test s'ha fet concatenant seqüències del *dataset* Weizmann amb seqüències del *dataset* Hollywood i mesurant els *scores* de les dues classes, de forma similar a l'experiment anterior. La detecció s'ha fet seguint el mètode primer del experiment anterior, és a dir, mesurant quina SVM dóna un *score* més alt entre dos mínims locals, i la classe corresponent a aquesta SVM serà la que correspon a la classe reconeguda. La figura 53 mostra un exemple de la resposta de les dues SVMs a la seqüència d'accions concatenades d'entrada.

El resultat ha estat del 100% de *recognition rate*. Malauradament, no tenim constància de l'ús d'aquest mètode d'avaluació per altres autors i per tant no poden fer una comparativa.

S'ha de dir que els context de les accions del *dataset* Hollywood són molt diferents de les del *dataset* Weizmann i que això pot haver afavorit els bons resultats aconseguits en aquest experiment, però considerem que tot i així, l'experiment continua tenint valor.

Figura 53. Mostra de la resposta de les dues SVMs a una seqüència completa del *dataset* Weizmann concatenada amb accions del *dataset* Hollywood.

A.5.6 Conclusions

En aquest treball hem presentat una tècnica de reconeixement d'accions que és capaç de reconèixer accions concatenades a una seqüència de vídeo d'entrada.

A diferencia de altres tècniques basades en finestra lliscant, nosaltres no apliquem el reconeixement directament sobre els *frames* d'entrada, sinó que ho fem sobre la superfície generada per la projecció d'una variant de la transformada *R* aplicada sobre el OF de les imatges d'entrada.

Hem aplicat una tècnica de detecció de màxims i mínims locals sobre la informació dels *scores* de SVMs entrenades per a cada classe per detectar els inicis i finals de les accions.

Els resultat experimentals han demostrat que la nostra tècnica obté molt bon rendiment en el reconeixement de les 10 accions concatenades del *dataset* públic Weizmann arribant a un 95,5% de *recognition rate*.

El resultats demostren també que la tècnica aconsegueix un 100% de *recognition rate* quan l'entrenem per reconèixer entre accions conegudes i no conegudes.

B. Llistat de publicacions

Aquesta tesi ha generat les següents publicacions:

Journal Papers

Josep Maria Carmona, Joan Climent. Human action recognition by means of subtensor projections and dense trajectories. Pattern Recognition Journal, vol 81, pp. 443-455, April 2018. https://doi.org/10.1016/j.patcog.2018.04.015.

Conference Contributions

Josep Maria Carmona, Joan Climent. Temporal Segmentation of Human Actions in Video Sequences. SAI Intelligent Systems Conference 2017 (IntelliSys 2017), London, UK, September 2017.

Josep Maria Carmona, Joan Climent. Action recognition using the *Rf* Transform on optical flow images. International Conference on Computer Vision Theory and Applications, p. 266-271, Porto, Portugal, 2017.

Josep Maria Carmona, Joan Climent. A performance evaluation of HMM and DTW for gesture recognition. Iberoamerican Congress on Pattern Recognition p. 236-243, Buenos Aires, Argentina, 2012.

Referències

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

[2] N. T. Nguyen, D. Q. Phung, S. Venkatesh, and H. Bui, “Learning and detecting activities from movement trajectories using the Hierarchical Hidden Markov Models”, CVPR, 2, 955-960, 2005.

[3] Daniel Weinland, Remi Ronfard, Edmond Boyer, A survey of vision-based methods for action representation, segmentation and recognition, In Computer Vision and Image Understanding, Volume 115, Issue 2, 2011, Pages 224-241.

[4] M. S. Ryoo and J. K. Aggarwal, “Recognition of composite human activities through context-free grammar based representation”, CVPR, 1709-1718, June 2006.

[5] Klaeser, A., Marszalek, M., Schmid, C. “A spatio-temporal descriptor based on 3d-gradients”, BMVC, pp. 995-1004, 2008.

[6] Scovanner, P., Ali, S., Shah, M. “A 3-dimensional sift descriptor and its application to action recognition”, ACM Multimedia, pp. 357-360, 2007.

[7] Zhang, Z., Hu, Y., Chan, S., Chia, L. “Motion context: A new representation for human action recognition”. ECCV, Part IV. LNCS, vol. 5305, pp. 817-829, 2008.

[8] Serre, T.; Wolf, L. & Poggio, T., “Object Recognition with Features Inspired by Visual Cortex”., in 'CVPR (2)' , IEEE Computer Society, , pp. 994-1000 , 2005.

[9] T. Serre, L. Wolf, S. Bileschi, M. Riesenhuber, T. Poggio. “Robust object recognition with cortex-like mechanisms”, IEEE Transactions on Pattern Analysis and Machine Intelligence 29, 411-429, 2007.

[10] A. Bosch, A. Zisserman, and X. Munoz. “Image classification using random forests and ferns”. In International Conference on Computer Vision, 2007.

[11] Lowe, D.: “Distinctive image features from scale-invariant keypoints”. IJCV 20, 91-110, 2003.

[12] M. Leo, T. D’Orazio, I. Gnoni, P. Spagnolo, and A. Distante, “Complex human activity recognition for monitoring wide outdoor environments”, ICPR, 4, 913-916, Aug. 2004.

[13] Zhenhua Wang, B. Fan and F. Wu, "Local Intensity Order Pattern for feature description," *Computer Vision (ICCV), 2011 IEEE International Conference on*, Barcelona, pp. 603-610, 2011.

[14] H Jhuang, T Serre, L Wolf, T Poggio. “A biologically inspired system for action recognition Computer Vision”, IEEE 11th International Conference on, 1-8, 2007.

[15] Y. M. Lui, J. R. Beveridge, and M. Kirby. “Action classification on product manifolds”, In IEEE Conference on Computer Vision and Pattern Recognition, San Francisco, 2010.

[16] Kim and R. Cipolla. “Canonical correlation analysis of video volume tensors for action categorization and detection”, IEEE Transactions on Pattern Analysis and Machine Intelligence, 31(8):1415-1428, 2009.

[17] Lowe, D.: “Distinctive image features from scale-invariant keypoints”. IJCV 20, 91-110, 2003.

[18] N. Dalal and B. Triggs. Histograms of Oriented Gradients for Human Detection. In CVPR, pages 886-893, 2005

[19] Zhenhua Wang, B. Fan and F. Wu, "Local Intensity Order Pattern for feature description," *Computer Vision (ICCV), 2011 IEEE International Conference on*, Barcelona, 2011, pp. 603-610. doi: 10.1109/ICCV.2011.6126294

[20] A. Bosch, A. Zisserman, and X. Munoz. "Image classification using random forests and ferns". In International Conference on Computer Vision, 2007.

[21] J. B. Tenenbaum, V. de Silva, and J. C. Langford. A global geometric framework for nonlinear dimensionality reduction. Science, pages 2319{2323, 2000.

[22] S. T. Roweis and L. K. Saul. Nonlinear dimensionality reduction by locally linear embedding. Science, pages 2323{2326, 2000.

[23] G. Hinton and S. Roweis. Stochastic neighbor embedding. In Advances in Neural Information Processing Systems 15, pages 833{840, 2003.

[24] Bob de Graaf, Automatic hand gesture recognition using manifold learning. Thesis

[25] R. Souvenir and J. Babbs, "Learning the viewpoint manifold for action recognition," *2008 IEEE Conference on Computer Vision and Pattern Recognition*, Anchorage, AK, pp.1-7, 2008.

[26] Yui Man Lui and Ross Beveridge , "Tangent Bundle for Human Action Recognition", IEEE International Conference on Automatic Face and Gesture Recognition, Santa Barbara, 2011.

[27] S. Tabbone, L.Wendling, and J.-P. Salmon. A new shape descriptor defined on the radon transform. Comput. Vis. Image Underst., 102(1):42–51, 2006.

[28] Y. Sheikh, M. Sheikh, and M. Shah. Exploring the space of a human action. Proc. International Conference on Computer Vision, 1:144–149, 2005.

[29] Kim, T.-K. & Cipolla, R., "Gesture Recognition Under Small Sample Size", in Yasushi Yagi; Sing Bing Kang; In-So Kweon & Hongbin Zha, ed., 'ACCV (1)' , Springer, , pp. 335-344, 2007.

[30] A. Veeraraghavan, A. K. Roy-Chowdhury, and R. Chellappa. Matching shape sequences in video with applications in human movement analysis. IEEE Transactions on Pattern Analysis and Machine Intelligence, (12):1896–1909, 2005.

[31] Harandi, M. T.; Sanderson, C.; Wiliem, A. & Lovell, B. C., "Kernel analysis over Riemannian manifolds for visual recognition of actions, pedestrians and textures", in 'WACV' , IEEE Computer Society, , pp. 433-439, 2012.

[32] Rizwan Chaudhry, Avinash Ravichandran, Gregory D. Hager, René Vidal. "Histograms of oriented optical flow and Binet-Cauchy kernels on nonlinear dynamical systems for the recognition of human actions". CVPR, pp. 1932-1939, 2009.

[33] Yuan, Y.; Zheng, H.; Li, Z. & Zhang, D., "Video action recognition with spatio-temporal graph embedding and spline modelling", in 'ICASSP' , IEEE, , pp. 2422-2425, 2010.

[34] Wang H, Klaser A, Schmid C, Liu CL, Action recognition by dense trajectories. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 3169–3176, 2011.

[35] N. Dalal, B. Triggs, and C. Schmid. Human detection using oriented histograms of flow and appearance. In ECCV, 2006.

[36] Peng X, Wang L, Wang X, Qiao Y, Bag of visual words and fusion methods for action recognition: Comprehensive study and good practice, Computer Vision and Image Understanding, pp. 109 – 125, 2016.

[37] Zhang, J. , Marszalek, M. , Lazebnik, S. , Schmid, C. , Local features and kernels for classification of texture and object categories: a comprehensive study. Int. J. Comput. Vis. 73 (2), 213–238, 2007.

[38] Lingqiao Liu, Lei Wang, and Xinwang Liu, In defense of soft-assignment coding. In *Proceedings of the 2011 International Conference on Computer Vision* (ICCV '11). IEEE Computer Society, Washington, DC, USA, 2486-2493, 2011.

[39] Wang, J. , Yang, J. , Yu, K. , Lv, F. , Huang, T.S. , Gong, Y. , Locality-constrained linear coding for image classification. In: CVPR, pp. 3360–3367, 2010.

[40] Jégou, H. , Perronnin, F. , Douze, M. , Sánchez, J. , Pérez, P. , Schmid, C.. Aggregat- ing local image descriptors into compact codes. IEEE Trans. Pattern Anal. Mach. Intell. 34 (9), 1704–1716, 2012.

[41] Zhou, X. , Yu, K. , Zhang, T. , Huang, T.S. , Image classification using super-vector coding of local image descriptors. In: ECCV, pp. 141–154, 2010.

[42] A. F. Bobick and Y. A. Ivanov. Action Recognition Using Probabilistic Parsing. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (CVPR '98). IEEE Computer Society, Washington, DC, USA, 196, 1998.

[43] Peng, X. , Wang, L. , Cai, Z. , Qiao, Y. , Peng, Q. , Hybrid super vector with improved dense trajectories for action recognition. ICCV Workshop on Action Recognition with a Large Number of Classes, 2013

[44] Wu, J. , Zhang, Y. , Lin, W, Towards good practices for action video encoding. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2577-2584, 2014

[45] Hyun-Joo Jung, Ki-Sang Hong, Modeling temporal structure of complex actions using Bag-of-Sequencelets, In Pattern Recognition Letters, Volume 85, Pages 21-28, 2017.

[47] M. A. R. Ahad, T. Ogata, J. K. Tan, H. S. Kim and S. Ishikawa, "Motion recognition approach to solve overwriting in complex actions," *2008 8th IEEE International Conference on Automatic Face & Gesture Recognition*, Amsterdam, pp. 1-6, 2008.

[48] A. D. Wilson and A. F. Bobick, "Parametric hidden Markov models for gesture recognition," in *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 21, no. 9, pp. 884-900, Sep 1999.

[49] Tian-Shu Wang, Heung-Yeung Shum, Ying-Qing Xu, and Nan-Ning Zheng. Unsupervised Analysis of Human Gestures. In *Proceedings of the Second IEEE Pacific Rim Conference on Multimedia: Advances in Multimedia Information Processing* (PCM '01), Heung-Yeung Shum, Mark Liao, and Shih-Fu Chang (Eds.). Springer-Verlag, London, UK, UK, 174-181, 2001.

[50] View invariant identification of pose sequences for action recognition Abhijit S. Ogale, Alap Karapurkar, Gutemberg Guerra-Filho and Yiannis Aloimonos Computer Vision Laboratory, University of Maryland, College Park, MD 20742, 2004.

[51] L. Zelnik-Manor and M. Irani, "Event-based analysis of video," *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001*, pp. II-123-II-130 vol.2, 2001.

[52] M. Brand and V. Kettnaker, "Discovery and segmentation of activities in video," in *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 8, pp. 844-851, 2000.

[53] P. Peursum, H. H. Bui, S. Venkatesh and G. West, "Human action segmentation via controlled use of missing data in HMMs," *Proceedings of the 17th International Conference on Pattern Recognition, 2004. ICPR 2004.*, pp. 440-445 Vol.4, 2004.

[54] L. R. Rabiner. A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286, February 1989.

[55] Cristian Sminchisescu, Atul Kanaujia, Dimitris Metaxas, Conditional models for contextual human motion recognition, In Computer Vision and Image Understanding, Volume 104, Issues 2–3, Pages 210-220, 2006.

[56] J. Lafferty, A. McCallum, F. Pereira, Conditional random fields: probabilistic models for segmenting and labeling sequence data, in: Internat. Conf. on Machine Learning, 2001.

[57] A. McCallum, D. Freitag, F. Pereira, Maximum entropy Markov models for information extraction and segmentation, in: Internat. Conf. on Machine Learning, 2000.

[58] R. Filipovych and E. Ribeiro, "Learning human motion models from unsegmented videos," *2008 IEEE Conference on Computer Vision and Pattern Recognition*, Anchorage, AK, pp. 1-7, 2008.

[59] Ling Shao, Ling Ji, Yan Liu, Jianguo Zhang, Human action segmentation and recognition via motion and shape analysis, In Pattern Recognition Letters, Volume 33, Issue 4, Pages 438-445, 2012.

[60] Sean Ryan Fanello, Ilaria Gori, Giorgio Metta, and Francesca Odone. Keep it simple and sparse: real-time action recognition. *J. Mach. Learn. Res.* 14, 1 (January 2013), 2617-2640, 2013.

[61] M. Hoai, Z. Z. Lan and F. De la Torre, "Joint segmentation and classification of human actions in video," *CVPR 2011*, Providence, RI, pp. 3265-3272, 2011.

[62] Y. Ke, R. Sukthankar, and M. Hebert, “Event detection in crowded videos,” in IEEE International Conference on Computer Vision, October 2007.

[63] E. Shechtman and M. Irani, "Space-time behavior based correlation," *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, pp. 405-412 vol. 1, 2005.

[64] Yan Ke, R. Sukthankar and M. Hebert, "Efficient visual event detection using volumetric features," *Tenth IEEE International Conference on Computer Vision (ICCV'05) Volume 1*, pp. 166-173 Vol. 1, 2005.

[65] D. Weinland, R. Ronfard and E. Boyer, "Automatic Discovery of Action Taxonomies from Multiple Views," *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, pp. 1639-1645, 2006.

[66] D.Weinland, R. Ronfard, and E. Boyer. Motion history volumes for free viewpoint action recognition. In *PHI*, 2005.

[67] Hua Zhong, Jianbo Shi and M. Visontai, "Detecting unusual activity in video," *Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (CVPR 2004), pp. II-819-II-826 Vol.2, 2004.

[61] Cen Rao, Alper Yilmaz, and Mubarak Shah. 2002. View-Invariant Representation and Recognition of Actions. *Int. J. Comput. Vision* 50, 2, 203-226, 2002.

[62] Kolda, T. G., & Bader, B. W. "Tensor decompositions and applications", *SIAM Review, 51*, 455-500, 2009.

[63] Kiers, H. A. L. "Towards a standardized notation and terminology in multiway analysis", Journal of Chemometrics, 14 , pp. 105-122, 2000.

[64] K. Mikolajcyk and C. Schmid, An affine invariant interest point detector. ICCV, vol. 2350, pp. 128-142, 2002.

[65] C. Silpa-Anan and R. Hartley, "Optimised KD-Trees for Fast Image Descriptor Matching," Proc. IEEE Conf. Computer Vision and Pattern Recognition CVPR, 2008.

[66] Sequential forward selection (SFS for short) proposed by Whitney in 1971 (Whitney, A.W., A direct method of nonparametric measurement selection.IEEE Trans. Comput. 20, 1100–1103, 1971.)

[68] Zhe Lin, Zhuolin Jiang and L. S. Davis, "Recognizing actions by shape-motion prototype trees," *2009 IEEE 12th International Conference on Computer Vision*, Kyoto, pp. 444-451, 2009.

[69] Yui Man Lui. "Human gesture recognition on product manifolds", Journal of Machine Learning Research 13, 3297-3321, 2012.

[70] M. Budnik, E. L. Gutierrez-Gomez, B. Safadi and G. Quénot, "Learned features versus engineered features for semantic video indexing," 2015 13th International Workshop on Content-Based Multimedia Indexing (CBMI), Prague, pp. 1-6, 2015.

[71] Z.-Z. Lan, M. Lin, X. Li, A. G. Hauptmann, and B. Raj., Beyond Gaussian pyramid: Multi-skip feature stacking for action recognition, CVPR, 2015.

[72] B. Fernando, E. Gavves, J. Oramas, A. Ghodrati, and T. Tuytelaars. Modeling video evolution for action recognition, CVPR, 2015.

[73] M. Hoai and A. Zisserman. Improving human action recognition using score distribution and ranking. In Proc. ACCV, 2014.

[74] Peng, X. , Zou, C. , Qiao, Y. , Peng, Q. , Action recognition with stacked fisher vectors, ECCV. Springer, pp. 581–595, 2014.

[75] Radon, J. U¨ ber die Bestimmung von Funktionen durch ihre Integralwerte l¨angs gewisser Mannigfaltigkeiten. Akad. Wiss., 69:262–277, 1917.

[76] J. Y. Ng, M. J. Hausknecht, S. Vijayanarasimhan, O. Vinyals, R. Monga, and G. Toderici. Beyond short snippets: Deep networks for video classification. In Proc. CVPR, 2015.

[77] K. Simonyan and A. Zisserman, Two-stream convolutional networks for action recognition in videos, NIPS, 2014.

[78] H. Bilen, B. Fernando, E. Gavves, A. Vedaldi, S. Gould, Dynamic Image Networks for Action Recognition, IEEE Conference on Computer Vision and Pattern Recognition 2016.

[79] H. Wang and C. Schmid. Action recognition with improved trajectories. In Proc. ICCV, pages 3551–3558, 2013.

[80] Pavlovic, V.I., Sharma, R., Huang, T.S.: Visual interpretation of hand gestures for human

computer interaction. IEEE Trans. Pattern Anal. Mach. Intell. 19(7), 677–695, 1997.

[81] Wexelblat, A.: An approach to natural gesture in virtual environments. ACM Transactions on Computer-Human Interaction 2(3), 179–200, 1995.

[82] Sakoe, H., Chiba, S.: Dynamic programming algorithm optimization for spoken word recognition. IEEE Trans. Acoustics, Speech, and Signal Processing 26(1), 43–49, 1978.

[83] Lee, H., Kim, J.: An HMM-based threshold model approach for gesture recognition. IEEE Trans. Pattern Analysis and Machine Intelligence 21(10), 961–973, 1999.

[84] Wilson, A.D., Bobick, A.F.: Parametric hidden Markov models for gesture recognition. IEEE Trans. Pattern Analysis and Machine Intelligence 21(9), 884–900, 1999.

[85] J. M. Carmona and J. Climent, "A Performance Evaluation of HMM and DTW for Gesture Recognition," Proceedings of the 17th Iberoamerican Congress on Pattern Recognition (CIARP), Buenos Aires, Argentina, pp. 236-243, 2012.

[86] Csurka, G.; Bray, C.; Dance, C. & Fan, L., "Visual categorization with bags of keypoints", Workshop on Statistical Learning in Computer Vision, ECCV , 1-22, 2004.

[87] Niebles, J. C.; Wang, H. & Li, F.-F. "Unsupervised Learning of Human Action Categories Using Spatial-Temporal Words", International Journal of Computer Vision 79 (3) , 299-318 , 2008.

[88] M. Giese and T. Poggio. Neural mechanisms for the recognition of biological movements and action. Nat. Rev. Neurosci.,4:179–192, 2003.

[89] J. Niebles and L. Fei-Fei. A hierarchical model of shape and appearance for human action classification. In CVPR, 2007.

[90] Y. Wang, K. Huang and T. Tan, "Human Activity Recognition Based on R Transform," *2007 IEEE Conference on Computer Vision and Pattern Recognition*, Minneapolis, MN, pp. 1-8, 2007.

[91] Weiming Hu, Li Li, Qingdi Wei, Human Activity Recognition Based on R Transform and Fourier Mellin Transform. ISVC (2) : 631-640, 2009.

[92] Karlsson, S. M. & Bigün, J. Lip-motion events analysis and lip segmentation using optical flow, in 'CVPR Workshops' , IEEE, pp. 138-145, 2012.

[93] Gorelick, L.; Blank, M.; Shechtman, E.; Irani, M. & Basri, R. “Actions as Space-Time Shapes”, IEEE Trans. Pattern Anal. Mach. Intell. 29 (12) , 2247-2253, 2007.

[94] C. Sch¨uldt, I. Laptev, and B. Caputo. Recognizing human actions: A local svm approach. In ICPR, Cambridge, UK, 2004.

[95] K. Soomro, A. R. Zamir, and M. Shah. UCF101: A dataset of 101 human actions classes from videos in the wild. CoRR, abs/1212.0402, 2012.

[96] H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, and T. Serre. HMDB: A large video database for human motion recognition. In Proc. ICCV, pages 2556–2563, 2011.

[97] D.H. Hubel and T.N. Wiesel, “Receptive Fields, Binocular Interaction and Functional Architecture in the Cat’s Visual Cortex,” J. Physiology, vol. 160, pp. 106-154, 1962.

[98] L. Maaten and G. Hinton, Visualizing data using t-sne, Journal of Machine Learning Research, 2008.

[99] J. B. Tenenbaum, V. de Silva, and J. C. Langford. A global geometric framework for nonlinear dimensionality reduction. Science, pages 2319{ 2323, 2000.

[100] Lazebnik, S.; Schmid, C. & Ponce, J., “Beyond Bags of Features: Spatial Pyramid Matching for Recognizing Natural Scene Categories”, in 'CVPR (2)' , IEEE Computer Society, , pp. 2169-2178, 2006.

[101] Laptev, I. & Lindeberg, T., “Space-time interest points”, in 'Computer Vision, 2003. Proceedings. Ninth IEEE International Conference on' , pp. 432-439, 2003.

[102] O'Hara, S. & Draper, B. A. “Introduction to the Bag of Features Paradigm for Image Classification and Retrieval”,CoRR abs/1101.3354, 2011.

[103] Arodz, T. Invariant object recognition using radonbased transform. Computers and Artificial Intelligence, 24:183–199, 2005.

[104] S. Ubalde, N. A. Goussies, and M. E. Mejail. 2014. Efficient descriptor tree growing for fast action recognition. *Pattern Recogn. Lett.* 36, 213-220, 2014.

[105] Y. M. Lui. “Tangent bundles on special manifolds for action recognition”, IEEE Transactions on Circuits and Systems for Video Technology, 22(6):930-942, 2012b.

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

[107] Kim, T.K.,Wong, S.F., Cipolla, R.: Tensor canonical correlation analysis for action classification. In: In Proc. of IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2007).

[108] Cunzhao Shi, Yanna Wang, Fuxi Jia, Kun He, Chunheng Wang, Baihua Xiao, Fisher vector for scene character recognition: A comprehensive evaluation, In Pattern Recognition, Volume 72, Pages 1-14, 2017.

[109] Meng, H; Pears, N; Freeman, M; Bailey, C; Motion history histograms for human action recognition. In: Kisačanin, B and Bhattacharyya, SS and Chai, S, (eds.) Embedded Computer Vision. (pp. 139-162), 2008.

[110] Earnest Paul Ijjina, Krishna Mohan Chalavadi, Human action recognition in RGB-D videos using motion sequence information and deep learning, In Pattern Recognition, Volume 72, Pages 504-516, 2017.

[111] D. Tao, X. Li, X. Wu and S. J. Maybank, "General Tensor Discriminant Analysis and Gabor Features for Gait Recognition," in *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, no. 10, pp. 1700-1715, Oct. 2007.

[112] Weifeng Liu, Xinghao Yang, Dapeng Tao, Jun Cheng, Yuanyan Tang, Multiview dimension reduction via Hessian multiset canonical correlations, In Information Fusion, Volume 41, Pages 119-128, 2018.

[113] T. R. Almaev and M. F. Valstar, "Local Gabor Binary Patterns from Three Orthogonal Planes for Automatic Facial Expression Recognition," *2013 Humaine Association Conference on Affective Computing and Intelligent Interaction*, Geneva, pp. 356-361, 2013.

[114] Varol G, Laptev I and Schmid C, "Long-term temporal convolutions for action recognition" arXiv preprint arXiv p 1604.04494, 2016

[115] Bilen, Hakan; Fernando, Basura; Gravves, Efstratios; Vedaldi, Andrea, "Action Recognition with Dynamic Image Networks", IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017.

[116] Rohit Girdhar, Deva Ramanan, Abhinav Gupta, Josef Sivic, Bryan Russell, "ActionVLAD: Learning spatio-temporal aggregation for action classification", IEEE Conference on Computer Vision and Pattern Recognition (CVPR),2017

[117] I. C. Duta, B. Ionescu, K. Aizawa and N. Sebe, "Spatio-Temporal Vector of Locally Max Pooled Features for Action Recognition in Videos," IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, 2017, pp. 3205-3214.

[118] Christoph Feichtenhofer, Axel Pinz, Richard P. Wildes, "Spatiotemporal Multiplier Networks for Video Action Recognition", IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017

[119] Y. Wang, M. Long, J. Wang and P. S. Yu, "Spatiotemporal Pyramid Network for Video Action Recognition", IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, 2017, pp. 2097-2106.

[120] I. C. Duta, J. R. R. Uijlings, T. A. Nguyen, K. Aizawa, A. G. Hauptmann, B. Ionescu, and N. Sebe. Histograms of motion gradients for real-time video classification, CBMI, 2016.