Detección de grupos conversacionales en escenarios de video-protección con aglomeraciones de personas

Elvis Ferrera Cedeño[1], Edel García Reyes[1] y Niusvel Acosta-Mendoza[1, 2]

[1] Centro Nacional de Aplicaciones de Tecnologías de Avanzadas (CENATAV), 7ª No. 21406 /e 214 y 216, Siboney Playa. CP: 12200, Habana, Cuba
[2] Instituto Nacional de Astrofísica. Óptica y Electrónica (INAOE) Luis Enrique Erro No. 1, Sta. María Tonantzintla, Puebla, México
elvis.ferrera@cenatav.co.cu,egarcia@cenatav.co.cu, nacosta@cenatav.co.cu

Resumen. Una de las tareas de la video-protección es entender el comportamiento de las escenas con aglomeraciones de personas. Esto garantiza la custodia y salvamento de las personas involucradas en dichas escenas mediante la detección de situaciones peligrosas (motines, manifestaciones, accidentes, etc.). Muchas investigaciones en Visión por Computadoras y el Reconocimiento de Patrones se han enfocado en la automatización de dicha tarea analizando las interacciones de los grupos de personas que conforman las escenas; siendo de interés en este trabajo los grupos conversacionales estacionarios. Basado en el concepto de F-Formación, que es una definición de grupo donde las relaciones entre las personas son dadas por patrones espaciales y de orientación, en este trabajo se propone un nuevo algoritmo para la detección de grupos conversacionales utilizando información temporal en los videos. El comportamiento del algoritmo propuesto es evaluado sobre una base de datos sacada de un entorno real. Los resultados experimentales muestran la efectividad de la propuesta.

Palabras claves: Video-protección; f-formación; aglomeración de persona; grupos conversacionales estacionarios

1 Introducción

Con el incremento de la población y la diversidad de las actividades humanas, son cada vez más frecuentes las escenas de aglomeraciones de personas (ejemplos en la figura 1). El análisis de estas escenas es importante para predecir y prevenir situaciones de peligros (accidentes, motines, manifestaciones, peleas callejeras, etc.). Dichas tareas son realizadas por la video-protección, encargada del monitoreo diario a distancia, como una de las técnicas no invasivas para velar por el cuidado de las personas y objetivos importantes en una nación.

La realización de esta actividad por los humanos es una tarea desafiante. Aunque tengan habilidades para aprender patrones de comportamientos en escenas complicadas y tomar una rápida decisión, estudios psicológicos [1] indican que dichas habilidades se ven afectadas cuando tienen que lidiar con más de una señal simultáneamente. En

las escenas de aglomeraciones de personas hay que velar por un gran número de individuos y diversidad de actividades.

(a) Densidad media (b) Densidad baja (c) Densidad alta

Figura 1. Aglomeraciones de personas

La Visión por Computadoras y el Reconocimiento de Patrones han mostrado gran interés por la modelación y automatización de esta actividad. El proceso es complejo debido a las constantes oclusiones y diversidad de actividades que presentan las escenas con aglomeraciones. Esto ha estado respaldado por estudios sociales, biológicos y psicológicos donde una nueva área ha surgido nombrada; procesamiento de señales sociales [2] donde se revela que las escenas de aglomeraciones están compuestas por pequeños grupos [3]. Se evidencia la importancia de la detección de grupos para el análisis del comportamiento. Algunos trabajos como [4] utilizan dicho enfoque, la presente investigación también lo adopta. Se propone un método de detección de grupos basado en el concepto de F-Formación sobre escenas de baja y media densidad (ver las figuras 1(a) ,1(b)). Cabe mencionar que existen otros tipos de grupos como en [5] pero quedan fuera del alcance de este trabajo. Las principales contribuciones se listan a continuación:

- Una función de pertenencia que describe las relaciones entre los individuos en un video.
- Un corte sobre una matriz de fusión.
- Método para estimar grupos basado en el concepto de F-Formación sobre secuencias de videos.

Este escrito sigue con la subsección 1.1 donde se define un conjunto de conceptos básicos. La sección 2 es dedicada a los trabajos relacionados. Luego en la sección 3 se describe el método propuesto. Los resultados experimentales son discutidos en la sección 4. Las conclusiones de este trabajo y algunas ideas hacia futuras investigaciones son plasmadas en la sección 5.

1.1 Definiciones básicas

Para la comprensión del presente artículo se brinda un conjunto de definiciones a continuación.

Definición 1. *Una F-Formación tiene lugar cuando dos o más personas sostienen una relación de orientación y espacial, donde el espacio formado entre ellos es accedido en igualdad de condiciones*.

Definición 2. *Sea* $X, Y \subseteq R$ *conjuntos universales, entonces* $\tilde{R} = \{((x,y), \mu_{\tilde{R}}(x,y)) | \ (x,y) \subseteq X \times Y \}$ *es nombrada relación difusa sobre* $X \times Y$ *donde* $\mu_{\tilde{R}}(x,y)$ *es la función de pertenencias.*

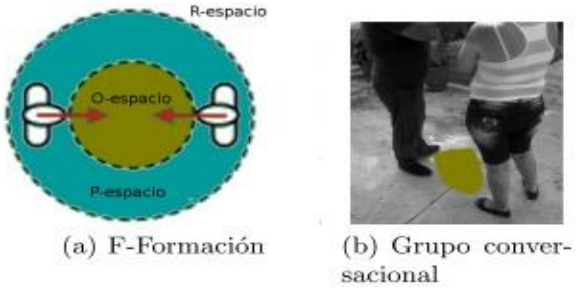
Definición 3. *Una matriz difusa* $\widetilde{M}$ *está compuesta por* $m \times n$ *relaciones difusas*

$$\widetilde{M} = \begin{bmatrix} \mu_{\tilde{R}}(x_1, y_1) & \cdots & \mu_{\tilde{R}}(x_1, y_n) \\ \vdots & \ddots & \vdots \\ \mu_{\tilde{R}}(x_m, y_1) & \cdots & \mu_{\tilde{R}}(x_m, y_n) \end{bmatrix}$$

Definición 4. *Sea* $\tilde{R}$ *una relación difusa, un* α*-corte es una relación* $R_\alpha = \{(x,y) | \ \mu_{\tilde{R}}(x,y) \geq \ \alpha, \ (x,y) \ \in \ \tilde{R} \}$.

Definición 5. *Una S-norma es una función* $S: [0,1]^2 \longrightarrow [0,1]^2$ *que satisface las propiedades siguientes:*

- *Conmutatividad:* $S(a,b) = S(b,a)$
- *Monótona:* $S(a,b) \leq S(c,d) \ si \ \ a \ \leq c \ \ y \ \ b \ \leq d$
- *Asociativa:* $S\big(a, S(b,c)\big) = S(S(a,b),c)$



(a) F-Formación (b) Grupo conversacional

Figura 2. Representación de una F-Formación

En [6] se plantea que en la práctica es mejor separar la F-Formación por tres espacios sociales (figura 2(a)): O-espacio, P-espacio y R-espacio. El O-espacio es un espacio

vacío que está rodeado de los individuos que se orientan hacia él. Es el más importante y su detección es concebida como una F-Formación. El P-espacio envuelve al O-espacio y a los cuerpos de las personas en la interacción, mientras que el R-espacio está más allá del P-espacio.

2. Trabajos relacionados

Dos enfoques han sido propuestos para descubrir F-Formaciones en imágenes. El primero, basado en la transformada de Hough [7]. Estos métodos realizan una estrategia de votos para crear un espacio acumulador y encontrar máximos locales. Los máximos representan el centro de cada O-espacio. Los individuos que forman parte de un grupo votarán por el mismo centro. En [6] cada persona puede realizar una cantidad fija de votos sobre una posición del plano tierra (x, y) mediante una distribución gaussiana. El voto k del individuo i tiene asociado un peso $w_{i,k}$ que representa la probabilidad de ser extraído de su distribución $w_{i,k} = N(s_{i.k}|\mu_1, \Sigma)$, los votos son acumulados en un espacio $Ac(x, y)$. Los máximos locales (O-espacios) sobre Ac se obtienen mediante $|(x, y)|.\sum w_{i,k}$, donde $|(x, y)|$ representa la cantidad de personas i que votaron por (x, y). Una mejora fue propuesta en [8] donde se detectan grupos de diferentes cardinalidades y los votos son agregados con una función de entropía (Boltzmann). En [15] se construye una matriz de patrones de relaciones por cada fotograma. Tres condiciones tienen que cumplirse para determinar si las personas i y j están relacionadas (primero, las personas tienen que estar en un rango de 2m, segundo, su foco de atención debe de solaparse y tercero, sus cabezas deben de estar orientadas mutuamente). La matriz de patrones de relaciones $M(i, j, t)$ toma valor 1 en cada posición si se cumplen las condiciones y 0 en otro caso.

El segundo enfoque se basa en la teoría de grafos. Las afinidades personales representan aristas en un grafo. Las F-Formaciones son representadas con cliques maximales. Para su detección se han usado, algoritmos de agrupamiento sobre grafos. En [9] las F-Formaciones son conjuntos dominantes, lo cual es una generalización de los cliques maximales en grafos con pesos. En [10,11] se desarrolla un esquema basado en la teoría de juegos. El agrupamiento es realizado sobre las estrategias definidas en la matriz de afinidad. El juego consiste en que cada jugador toma simultáneamente un elemento de la matriz recibiendo como pago el valor de la estrategia seleccionada. Se desarrolla en un entorno evolutivo donde solo un conjunto de puras estrategias (F-Formaciones) sobreviven. En [11] las matrices de afinidad obtenidas en cada fotograma son fusionadas utilizando un enfoque multiobjetivo para luego obtener los grupos. Una comparación entre los dos enfoques se realiza en [12]. Se muestra que los métodos basados en Hough son más efectivos utilizando posición y orientación mientras que los basados en grafos, con posiciones solamente. Se considera que la comparación es bastante simplista. No se tiene en cuenta el nivel de cómputo, siendo elevado en ambos enfoques. Los espacios de Hough crecen considerablemente con al aumento de los parámetros y sus valores (número de personas y cantidad de votos asignada) consumiendo gran cúmulo de memoria no siempre disponibles para estos casos. Por otro lado los

métodos basados en grafos pertenecen al conjunto de problemas duros. El problema de descubrir cliques maximales es NP-completo. El nivel de cómputo es una de las características que se debe de tener en cuenta para el despliegue en la práctica.

El método propuesto en esta investigación plantea un nuevo enfoque. Logrando un equilibrio entre eficiencia y eficacia.

3. Método propuesto

Esta sección describe el modelo propuesto para detectar grupos sociales en muchedumbres, como muestra la figura 3. Primero, basado en las posiciones y orientaciones obtenidas de cada persona en el video, se calcula el grado de relación entre ellas. Luego, se crea un conjunto de matrices difusas sobre una ventana de tiempo para fusionarlas. Finalmente, se realiza un **α**-corte para obtener los grupos (F-Formaciones).

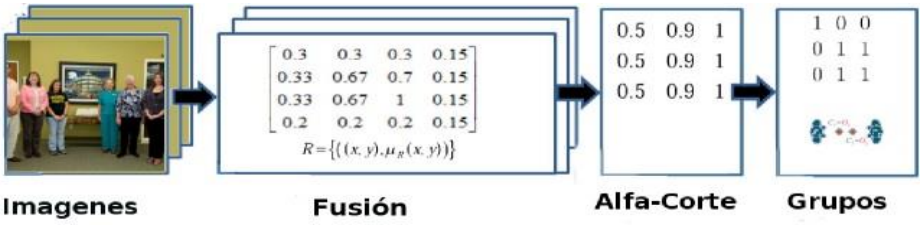


Figura 3. Flujo del modelo de detección de grupo

3.1 Grado relacional

Los grupos conversacionales (personas tomando café, jugando ajedrez, esperando en una cola, comiendo en un restaurante, etc.) mantienen las relaciones de sus miembros en el tiempo. Con un interés común[1], basado en el objetivo del grupo se mantienen la mayor parte del tiempo orientados hacia dicho objetivo. Teniendo en cuenta lo planteado se propone una función $\mu_{\tilde{R}}$ para medir el grado de relaciones entre los individuos, basada en las distancias entre ellos y sus objetivos.

$$\mu_{\tilde{R}} = \frac{1}{e^{\frac{d_1\theta_2 + d_2\theta_1}{\theta_1\theta_2}}} \qquad (1)$$

En la ecuación (1) $d_1 = \sqrt{(p_i - p_j)^2}$ donde p_i y p_j representan las posiciones sobre el plano de cada persona. También $d_2 = \sqrt{(c_i - c_j)^2}$ es la distancia euclidiana

[1] Lugar hacia donde mira con frecuencia el grupo. Es el objetivo del grupo

pero entre las posiciones de los objetivos c_i y c_j perseguidos por p_i y p_j. Las posiciones de los objetivos son obtenidas de la manera siguiente:

$$c_k = [x_k + r\cos\alpha_k, y_k + r\sin\alpha_k], k = i, j \quad (2)$$

Algo similar mostrado en la ecuación (2) se ha propuesto en [6], para encontrar F-Formaciones en un espacio de Hough. Esta investigación a diferencia de [6] no utiliza una función Gaussiana para los votos. Esto garantiza que cada persona vote solo una vez por su objetivo, lo cual es deseable debido a que se utiliza información temporal.

Los parámetros θ_1 y θ_2 representan las proximidades entre cada personas y sus objetivos respectivamente. Sus valores son basados en la teoría de Hall [13], donde se caracterizan las relaciones de los individuos por distancias.

3.2 Fusión y α-corte

Los patrones relacionales, aparecen frecuentemente en los grupos conversacionales estacionarios a lo largo del tiempo. La estabilidad es una de las propiedades de los grupos [4]. Este tipo de grupo presenta gran estabilidad. Permanecen casi estacionarios en el espacio físico que ocupan y cada miembro está fuertemente relacionado con sus semejantes.

El método propuesto descubre las relaciones que ocurren con mayor fortaleza en un intervalo de tiempo. Para lograr dicho objetivo se construye una matriz difusa M_k con $n \times n$ relaciones por cada instante k donde n es la cantidad de personas. Luego se fusionan las matrices en el intervalo $[k, k+1]$, donde se obtiene una matriz resultante M_r. Los elementos $M_k(i,j) = \frac{S(M_k(i,j),\dots,M_{k+1}(i,j))}{t}$, se calculan con la definición de frecuencia planteada en [14], S es la S-norma. Finalmente se aplica un **α**-corte sobre el conjunto de relaciones en M_r y se obtienen los grupos.

4. Experimentos y resultados

La efectividad del método propuesto se ha evaluado en la base de datos Coffe Break (CB). Representa un evento social de personas interactuando y disfrutando de un café. Las imágenes son tomadas con una cámara simple de resolución 1440 × 1080. Con un máximo de 14 individuos organizados en grupos de 2 y 3 personas. La orientación de las cabezas ha sido estimada considerando 4 configuraciones (frente, trasero, izquierda, derecha). La posición extraída del seguimiento fue proyectada sobre el plano de la tierra. Dos secuencias fueron anotadas por psicólogos utilizando varios cuestionarios. Para un total de 45 imágenes para las secuencia 1 y 75 imágenes para la secuencia 2.

Tabla 1. Comparación entre los métodos que extraen f-formación sobre la base de datos Coffe Break

Método	Precisión	Sensibilidad	F1
IRPM [15]	0.60	0.41	0.49
HFF [6]	0.82	0.83	0.82
R-GTCG [11]	0.86	0.88	0.87
Propuesta	**1**	0.82	**0.90**

Como métrica se ha adoptado la propuesta en [8], donde un grupo es correctamente detectado si al menos $\left\lceil \left(\frac{2}{3}\right).K \right\rceil$ de sus miembros son encontrados, siendo K la cantidad de individuos en el grupo. Para cada secuencia se ha estimado la precisión (p), sensibilidad (s) y F1 como se muestra en la expresión siguiente:

$$p = \frac{tp}{tp+fp}, s = \frac{tp}{tp+fn}, F1 = 2.\frac{p.s}{p+s}$$

Los experimentos fueron codificados con MATLAB, sobre el sistema operativo Windows 8 y ejecutados en una CPU Intel Core Dou 2,66 GHz, 2GB RAM. Se probaron diferentes longitudes de ventanas, los mejores resultados se consiguieron con todas las imágenes de cada secuencia (45, 75) respectivamente. Un **α**-corte = 0,07 fue empíricamente seleccionado. Se utilizó un r = 0,75 al igual que en [6] y $\theta_2 = 1{,}20$ como la distancia máxima entre relaciones sociales. La tabla 1 muestra la comparación del método propuesto contra los publicados en la literatura sobre las dos secuencias de la base de datos CB. Como muestran los resultados, el método propuesto supera en precisión y F1 a los métodos recientemente publicados en la literatura.

5. Conclusiones

Basado en el concepto de F-Formación que es una definición de grupo donde las relaciones entre las personas están dadas por patrones espaciales y de orientación. Un nuevo enfoque es propuesto para descubrir grupos conversacionales estacionarios en videos con aglomeraciones de personas. Se define una función de pertenencia que cuantifica el grado relacional de los individuos. Además utilizando información temporal se fusionan un conjunto de matrices difusas y se obtienen los grupos con un **α**-corte elegido de forma empírica. Los resultados obtenidos sobre la base de datos Coffe Break muestran la efectividad del enfoque, superando en precisión y F1 a los métodos reportados en la literatura. Futuras investigaciones estarán dirigidas a validar la propuesta en otras bases de datos. Además se trabajará con matrices que representen grafos difusos que se agruparán en una matriz fusionada. Esto evitará la presencia de **α**-cortes dependiente de la base de datos. También se buscará una longitud de ventana que sea propicia para varios escenarios y se modelará la estancia de los grupos en el espacio físico que ocupan, para de esta manera comprender el comportamiento de una escena.

Referencias

1. Li, T., Chang, H., Wang, M., Ni, B., Hong, R., Yan, S.: Crowded scene analysis: A survey. Circuits and Systems for Video Technology, IEEE Transactions on 25(3) (2015) 367–386
2. Vinciarelli, A., Pantic, M., Bourlard, H.: Social signal processing: Survey of an emerging domain. Image and Vision Computing 27(12) (2009) 1743–1759
3. Moussa¨ıd, M., Perozo, N., Garnier, S., Helbing, D., Theraulaz, G.: The walking behaviour of pedestrian social groups and its impact on crowd dynamics. PloS one 5(4) (2010) e10047
4. Shao, J., Loy, C., Wang, X.: Scene-independent group profiling in crowd. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2014) 2219–2226
5. Solera, F., Calderara, S., Cucchiara, R.: Socially constrained structural learning for groups detection in crowd. (2015)
6. Cristani, M., Bazzani, L., Paggetti, G., Fossati, A., Tosato, D., Del Bue, A., Menegaz, G., Murino, V.: Social interaction discovery by statistical analysis of f-formations. In: BMVC. Volume 2. (2011) 4
7. Mukhopadhyay, P., Chaudhuri, B.B.: A survey of hough transform. Pattern Recognition 48(3) (2015) 993–1010
8. Setti, F., Lanz, O., Ferrario, R., Murino, V., Cristani, M.: Multi-scale f-formation discovery for group detection. In: 2013 IEEE International Conference on Image Processing, IEEE (2013) 3547–3551
9. Hung, H., Kr¨ose, B.: Detecting f-formations as dominant sets. In: Proceedings of the 13th international conference on multimodal interfaces, ACM (2011) 231–238
10. Vascon, S., Mequanint, E.Z., Cristani, M., Hung, H., Pelillo, M., Murino, V.:A game-theoretic probabilistic approach for detecting conversational groups. In: Asian Conference on Computer Vision, Springer (2014) 658–675
11. Vascon, S., Mequanint, E.Z., Cristani, M., Hung, H., Pelillo, M., Murino, V.: Detecting conversational groups in images and sequences: A robust game-theoretic approach. Computer Vision and Image Understanding 143 (2016) 11–24
12. Setti, F., Hung, H., Cristani, M.: Group detection in still images by f-formation modeling: A comparative study. In: 2013 14th International Workshop on Image Analysis for Multimedia Interactive Services (WIAMIS), IEEE (2013) 1–4
13. Hall, E.T.: The hidden dimensiondoubleday. New York (1966)
14. Delgado, M., Gonz´alez, A.: A frequency model in a fuzzy environment. International Journal of Approximate Reasoning 11(2) (1994) 159–174
15. Farenzena, M., Tavano, A., Bazzani, L., Tosato, D., Paggetti, G., Menegaz, G.,Murino, V., Cristani, M.: Social interactions by visual focus of attention in a three dimensional environment. In: Workshop on Pattern Recognition and Artificial Intelligence for Human Behaviour Analysis (PRAI* HBA). Volume 1. (2009)