Miglioramento del sistema di annotazione WATSS con tecniche di Computer Vision

Lorenzo Cioni
lore.cioni@gmail.com

1. Introduzione

L'obiettivo di questo elaborato è di migliorare, anche tramite tecniche della *Computer Vision*, il sistema di annotazione web *WATSS*[2].

WATSS, abbreviazione per *Web Annotation Tool for Surveillance Scenarios*, è un sistema di annotazione web per la creazione di un groundtruth di scenari di sorveglianza. Il sistema consente infatti di annotare persone all'interno dei singoli frame di un video, assegnandogli una posizione (determinata tramite una *bounding box*), una identità (tramite *avatar*), la parte visibile e le orientazioni del corpo e dello sguardo. E' inoltre possibile associare più persone ad un medesimo gruppo e il punto di interesse presso il quale la persona si trova.

Uno degli obiettivi è quello di introdurre nel sistema un meccanismo di predizione delle annotazioni, andando a generare, a partire da una o più annotazioni consecutive di una stessa persona, una serie di *proposals* per i frames successivi. L'elaborazione dell'immagine a questi scopi viene effettuata tramite l'utilizzo di OpenCV, una libreria open source, nativa per C++, per la Computer Vision e l'Image Analysis.

Viene poi presentato uno studio e l'implementazione di un sistema che consente di sfruttare la geometria della scena per proporre delle annotazioni possibili.

Nelle sezioni successive viene presentata inizialmente un'analisi comparativa tra i vari sistemi di annotazione che vanno a costituire l'attuale stato dell'arte. Viene poi presentata la parte relativa alle migliorie apportate al sistema e le tecniche di Computer Vision utilizzate. Infine viene presentata un'analisi di usabilità a posteriori, mettendo in evidenza anche eventuali sviluppi futuri per l'applicazione.

2. Stato dell'arte

In questa sezione viene presentata un'analisi comparativa di alcuni dei più famosi sistemi di annotazione esistenti. Lo scopo di questa ricerca è quello di individuare i le caratteristiche comuni ai vari strumenti e le loro limitazioni. Da questa analisi poi è stata stilata una lista di requisiti che portano WATSS ad essere in accordo con gli altri sistemi introducendo allo stesso tempo nuove caratteristiche.

L'analisi dei sistemi si è incentrata principalmente su 3 strumenti open source esistenti: *LabelMe*, *ViPER-GT* e *VATIC*. Per ciascuno dei sistemi è stata stilata una lista di caratteristiche offerte ed evidenziate le eventuali limitazioni. Infine è presentata un'analisi anche con il sistema WATSS.

2.1. LabelMe

LabelMe[3] è un sistema Web che consente l'annotazione di oggetti all'interno di immagini. Le singole annotazioni sono effettuate mediante la definizione di aree poligonali nell'immagine e l'assegnazione di una label. Il tool offre la possibilità di indicare se un oggetto annotato è occluso o meno da altri oggetti presenti nella scena (non consente però di individuare la parte occlusa o visibile).

Le annotazioni possono essere annidate, è possibile dunque etichettare oggetti che sono inclusi gli uni negli altri.

In aggiunta alle annotazioni di oggetti il sistema consente di annotare intere aree dell'immagine: questo è reso possibile andando inizialmente a delimitare una porzione di immagine ed associando ad essa una label. In questo caso l'area così definita viene colorata interamente ed è necessario stabilire se si tratta di un'area interna o esterna.

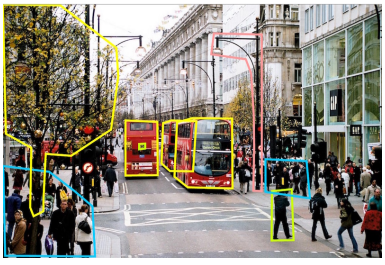

Figura 1. Annotazione di un'immagine tramite LabelMe

In fase di esportazione delle annotazioni viene generata una struttura in formato XML che è possibile importare nuovamente in un'altra immagine.

Il sistema non prevede la possibilità di generare *proposals* per le annotazioni, tutto il lavoro è a carico dell'utente.

2.2. ViPER-GT

ViPER-GT[4], acronimo di *Video Performance Evaluation resource*, è un sistema di annotazione per video e la generazione di un *groundtruth*.

Il sistema consente di annotare un video indicando cosa è contenuto nella scena, definendo un insieme di *classi* per ciascuna tipologia di contenuto. L'annotazione avviene manualmente da parte dell'utente che definisce delle *bounding boxes*, regioni di interesse dell'immagine, andando ad associare a ciascuna di esse una classe di appartenenza ed alcune metadati aggiuntivi, come ad esempio titolo, dimensione, etc. Le regioni possono essere di forme diverse: cerchi, ellissi, rettangoli e generici poligoni (definiti dai loro vertici).

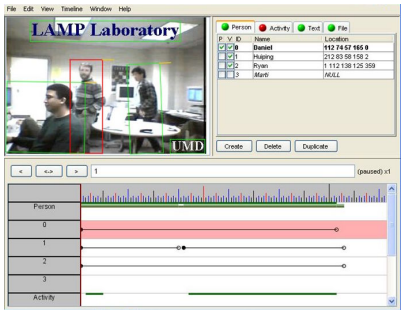

Figura 2. Interfaccia utente del tool ViPER-GT

Le annotazioni effettuate vengono visualizzate in una *timeline*: allo scorrere dei frame le annotazioni presenti nella scena corrente vengono evidenziate.

Il tool mette a disposizione anche un sistema di predizione delle annotazioni inserite basato sull'interpolazione lineare di più frame consecutivi. Questo metodo risulta molto efficace se si fornisce un numero di annotazioni, chiamate *ancore*, adeguato; con poche ancore definite la predizione risulta essere molto approssimativa.

2.3. VATIC

VATIC è un software di annotazione di video distribuito ai fini della ricerca nell'ambito della Computer Vision che consente la creazione di grandi dataset video. Il tool utilizza il sistema di crowdsourcing *Mechanical Turk* di Amazon.

Il sistema consente l'inserimento manuale di annotazioni per ciascun frame del video, definite mediante delle bounding box rettangolari.

Il tool dispone di una serie di plugin aggiuntivi che ne aumentano le potenzialità:

- *Tracking integration* per il tracciamento di oggetti in movimento nella scena
- *Sentence annotation* per l'annotazione di frasi e parole
- *Labeling time intervals* per l'annotazione di intervalli temporali
- *Human action labeling* per l'annotazione di azioni umane nella scena

Figura 3. Interfaccia utente del tool VATIC

Il tool è pensato principalmente per l'object detection nelle scene.

2.4. WATSS

WATSS[2], *Web Annotation Tool for Surveillance Scenarios*, è un sistema web per l'annotazione di dataset. Il tool è stato sviluppato per consentire l'annotazione del dataset *MuseumVisitors*[1] come parte del progetto MNEMOSYNE e rilasciato successivamente open source.

Il tool permette l'annotazione di persone e oggetti nella scena mediante la definizione di bounding boxes. In caso di occlusione è possibile poi indicare poi, mediante la definizione di una seconda bounding box, la parte visibile della persona o oggetto annotati.

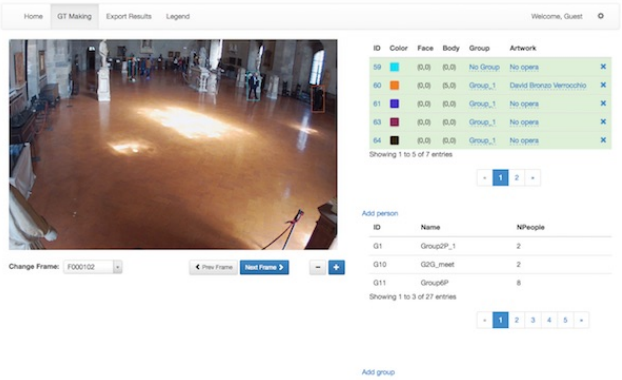

Figura 4. Interfaccia utente del tool WATSS

A ciascuna annotazione corrisponde un'*identità*: della persona annotata viene generato un avatar che consentirà di annotare la stessa persona negli altri frames del video. Per ciascuna annotazione è inoltre possibile indicare inoltre l'*orientazione del volto e del corpo* e il *punto di interesse* presso cui si trovano nella scena (nel caso del museo i punti di interesse sono rappresentati dalle varie opere d'arte).

In caso di presenza di gruppi di persone, è possibile indicare il gruppo di appartenenza definendo il nome dello

stesso, così da poterle poi riassociare in seguito anche nelle altre camere.

Il tool consente l'annotazione da parte di più utenti e la gestione di più camere.

3. Obiettivi

L'obiettivo principale del presente elaborato è di migliorare alcune caratteristiche del sistema di annotazione WATSS, utilizzando alcune tecniche di Computer Vision al fine di agevolare la creazione di annotazioni.

3.1. Interfaccia utente

Da un'analisi del sistema precedente, emergeva che alcune operazioni effettuabili tramite interfaccia risultavano essere poco intuitive per l'utilizzatore. In particolare, dai risultati del *System Usability Scale (SUS)* presentato in [2], si ritiene che non sia molto semplice imparare ad usare il sistema.

Le modifiche all'interfaccia grafica sono state dunque apportate con il fine di rendere più chiaro per l'utente le varie funzioni messe a disposizione dal sistema, a partire dalla schermata iniziale devono essere chiare fin da subito le sue caratteristiche e potenzialità.

3.2. Creazione e modifica delle annotazioni

La parte fondamentale del sistema è la fase di creazione e raffinamento delle bounding box all'interno della scena. Queste sono definite mediante una serie di rettangoli associati ad opportuni metadati.

Essendo dunque una fase fondamentale, deve essere semplice ed immediato per l'utente poter interagire con le annotazioni, modificandole e inserendole senza difficoltà. Il precedente sistema presenta alcune difficoltà, non consentendo una rapida modifica all'utente, rendendo l'azione di inserimento delle annotazioni leggermente complicata e difficile da gestire.

Obiettivo per questo aspetto è quello di introdurre un nuovo sistema di creazione e modifica delle annotazioni.

3.3. Timeline

Una delle caratteristiche mancanti in questo tool, presenti invece in molti altri sistemi di annotazione, è una *timeline*. Questa ha come scopo principale quello della navigazione tra i vari frames e la visualizzazione temporale delle annotazioni inserite.

Con questo strumento è infatti possibile visualizzare la durata di permanenza di una stessa persona in più frames consecutivi, consentendo all'utente di avere maggiore controllo sulla annotazioni inserite ed andare a correggere eventuali mancanze.

3.4. Predizione delle annotazioni

Dato il gran numero di frame da annotare, può risultare molto utile avere a disposizione un sistema di *predizione* delle annotazioni future in base ad una selezione corrente. Nel sistema è implementato un semplice meccanismo di predizione che ripropone una stessa bounding box nel frame successivo che può essere *approvata* con un click da parte dell'utente.

Mediante tecniche di Computer Vision si vuole fornire dei *proposals* per la posizione e la dimensione della stessa persona nei frame successivi. La predizione verrà valutata mediante la combinazione di più tecniche, come ad esempio la stima del moto, un *pedestrian detector* ed una stima mediante filtro di *Kalman*.

La fase di generazione dei proposals deve integrarsi nell'interfaccia, in particolar modo nella timeline.

3.4.1 Geometria della scena

Un altro tipo di predizione può essere effettuata inoltre conoscendo la geometria della scena. Questo è reso possibile se è nota la calibrazione delle telecamere con cui sono stati scattati i frames e se le telecamere sono fisse.

Utilizzando le informazioni spaziali è possibile ad esempio prevedere l'altezza di una persona data la sua posizione nella scena, consentendo così, ad esempio, di ridimensionare automaticamente la bounding box in base alla posizione in cui si vuole inserire.

4. Implementazione

4.1. L'interfaccia

L'interfaccia utente del sistema è stata rivisitata ed adattata alle nuove esigenze.

4.1.1 Pagina di creazione del groundtruth

La parte principale è costituita dal frame video su cui andremo ad aggiungere annotazioni e visualizzare quelle già esistenti. I due pannelli laterali per le persone e gruppi presenti nella scena sono stati organizzati in modo tale da consentirne una rapida navigazione ed uso. Come nella precedente versione dell'interfaccia, il pannello *People* mostra la lista delle annotazioni presenti nel frame visualizzato, ordinate in base all'identificativo delle persone. Per ciascuna persona viene mostrato il *gaze* del corpo e della faccia, il *gruppo di appartenenza* ed il *punto di interesse* presso cui si trovano.

Inserimento di una annotazione Tramite il pulsante *Add person* presente nel pannello delle annotazioni è possibile aggiungere una nuova annotazione al frame.

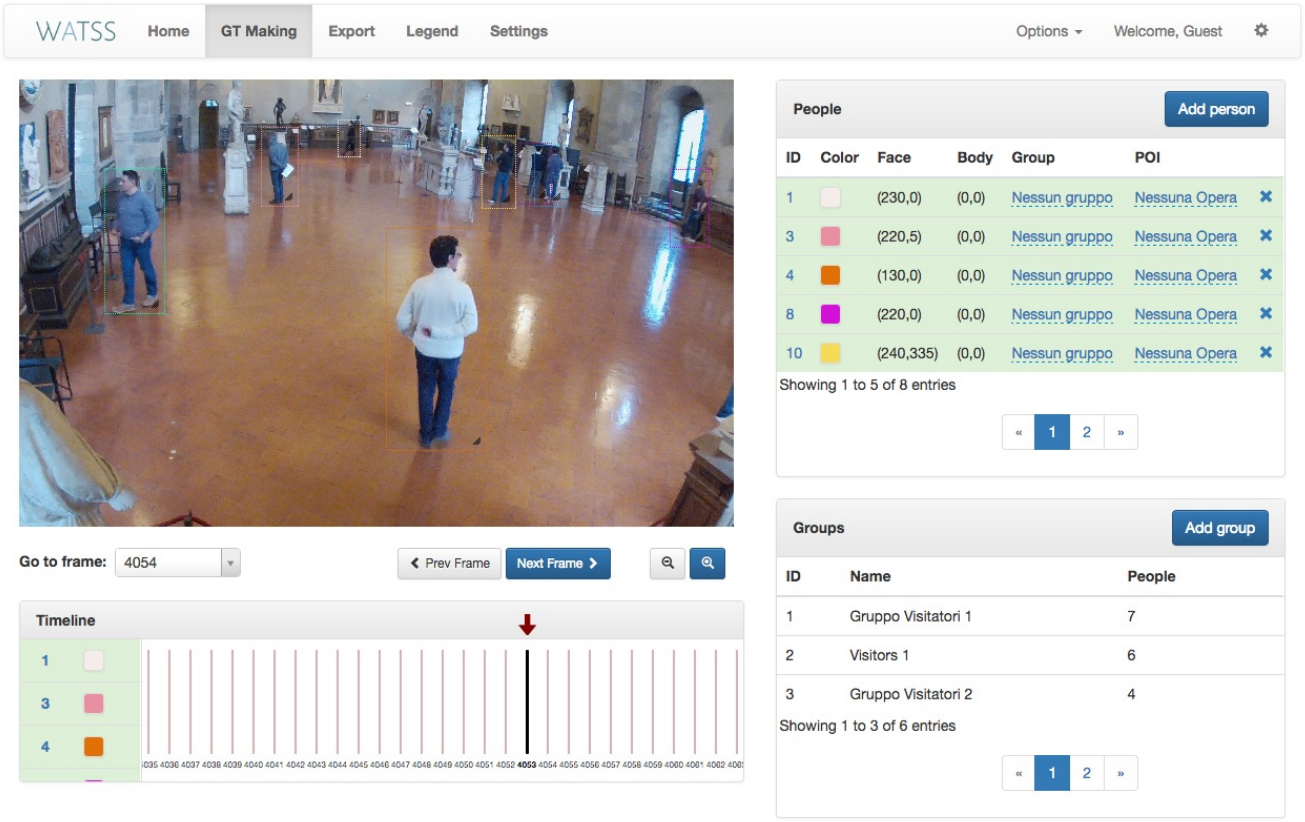

Figura 5. Interfaccia utente del tool WATSS

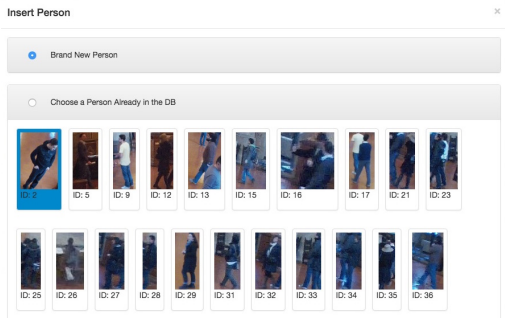

Figura 6. Aggiunta di una nuova annotazione

Come mostrato in Figura 6, in fase di creazione è possibile indicare se si vuole aggiungere un'annotazione rappresentante una nuova identità ancora non presente nel database oppure se si vuole aggiungere una nuova istanza di un'identità già presente e di cui viene mostrato un *avatar*.

Una volta selezionata l'opzione desiderata, la fase di creazione è diversa in base all'attivazione della *geometria della scena* o meno. Per l'attivazione e la disattivazione della geometria è sufficiente spuntare l'opzione presente nel menù *Options* della barra principale di navigazione.

In caso di geometria della scena *disattivata* la nuova annotazione verrà creata con la tecnica *click and drag*: l'utente clicca nel frame nel punto in cui vuole iniziare la sua selezione e tiene premuto spostando il mouse finchè la bounding box visualizzata non è della dimensione desiderata. A quel punto, rilasciando il click, la bounding box verrà inserita nella scena.

Se invece è attiva la geometria, questa verrà sfruttata per stimare l'altezza di una persona presente nella scena in base alla sua posizione nella stessa. La bounding box verrà automaticamente attaccata al puntatore del mouse e, muovendosi nella scena, sarà ridimensionata in base alla sua posizione. Una volta scelta la posizione, per effettuare l'inserimento sarà sufficiente effettuare un click nel punto desiderato.

In entrambi i casi, la procedura di inserimento può essere interrotta premendo il tasto *ESC*.

Modifica di una annotazione E' stata inoltre migliorata la fase di modifica delle annotazioni presenti. Le boun-

ding box sono ora trascinabili e ridimensionabili mediante il mouse.

Sia in fase di creazione che di modifica è ora possibile effettuare uno zoom del frame così da poter raffinare delle annotazioni. Questo si è reso necessario soprattutto per quanto riguarda oggetti e persone molto piccoli nella scena.

Selezionando una bounding box è infine possibile ridimensionarla semplicemente mediante la rotellina del mouse, consentendo una rapida scalatura del rettangolo definito.

La timeline La timeline è stata aggiunta all'interfaccia immediatamente sotto il frame. La timeline è così composta:

- *Elenco dei frames*: vengono visualizzati in ordine crescente in orizzontale. Viene visualizzato un sottoinsieme di frame adiacenti a quello corrente, evidenziato con un cursore a forma di freccia. Per ciascun frame viene indicata la presenza o meno di annotazioni (colore rosso) ed il numero.
- *Elenco delle annotazioni*: nel pannello laterale sinistro viene mostrato l'elenco delle annotazioni presenti nel frame corrente.

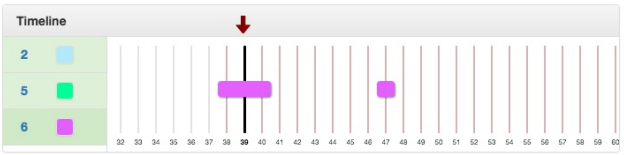

Figura 7. La *timeline*

E' possibile navigare tra i frame selezionandone uno dall'elenco e scorrerli mediante il trackpad o la rotellina del mouse.

Selezionando una persona nell'elenco viene mostrata la sua presenza nei frame della timeline mediante una barra orizzontale apposta sopra i frame in cui è presente la stessa persona.

A partire dalla barra visualizzata è possibile iniziare il processo di generazione dei *proposals* per i frame successivi. Per fare questo è sufficiente trascinare l'estremo destro della barra fino al frame entro il quale si vuole generare la predizione. Un esempio di generazione di proposals è mostrato in Figura 8.

4.1.2 Esportazione dei dati

La pagina di esportazione dei dati è stata ridisegnata al fine di consentire maggiori funzionalità. Nella precedente versione era possibile esportare un'unica tabella contenente le annotazioni presenti nei frames.

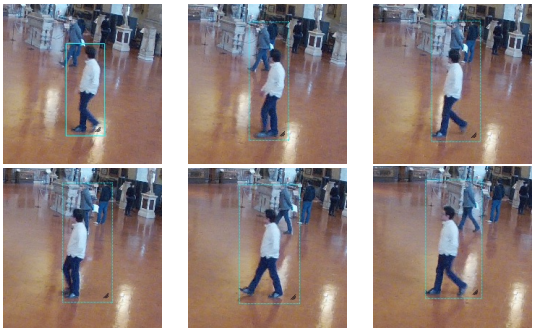

Figura 8. Generazione di *proposal* a partire da un'annotazione iniziale

Le funzioni di esportazione si suddividono ora in annotazioni e database: tramite il pannello *Annotations* è possibile selezionare i campi delle annotazioni desiderati e quali frame si vogliono estrarre, mentre tramite il pannello *Database* è possibile esportare la base dati (o solamente lo schema o la combinazione di schema e dati).

4.1.3 Configurazioni

Rispetto alla precedente versione, è stata creata una pagina di configurazione del sistema, al fine di rendere lo possibile l'interazione con gli elementi del database direttamente dall'interfaccia.

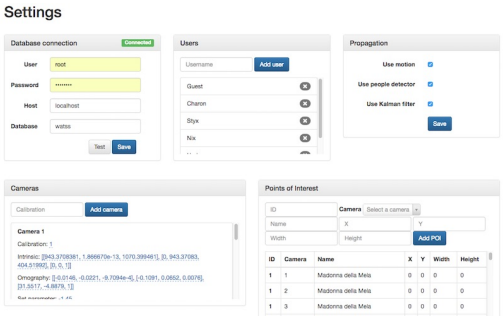

Figura 9. La pagina di configurazione

Nella pagina di configurazione è possibile:

- Configurare la *connessione al database*
- Inserire e rimuovere gli *utenti* che hanno accesso al sistema
- Aggiungere, rimuovere e impostare la *calibrazione* delle *camere*
- Aggiungere e rimuovere i *punti di interesse*
- Selezionare le tecniche usate nella generazione dei *proposals*

Per accedere alla pagina di configurazione è necessario aver preventivamente effettuato l'accesso al sistema.

4.1.4 Pagina di installazione

Per agevolare le operazioni di installazione è stata infine creata una pagina dedicata a questo scopo che guida nella procedura tramite interfaccia grafica.

La procedura di installazione viene descritta in seguito nella Sezione.

4.2. Tecniche utilizzate

Vengono ora presentate le tecniche e la teoria alla base rispettivamente della generazione dei *proposals* in fase di annotazione e della geometria della scena, usata nella creazione delle bounding box.

4.2.1 Generazione dei proposals

Il problema della generazione di *proposals* per frames successivi a quello/i annotato/i è affrontato mediante una combinazione di più tecniche per la stima del movimento di un oggetto nella scena, nel nostro caso di una persona.

Per questo scopo sono utilizzati:

- *Motion detection*: rilevazione del movimento mediante tecniche di rimozione dello sfondo di scena.
- *Pedestrian detector*: un detector di persone basato su *HOG features*.
- *Filtro di Kalman*: per la stima del moto data una serie di osservazioni passate.

Lo script di predizione prende in ingresso un insieme di frame con le rispettive annotazioni di una singola persona e l'insieme di frame su cui si vuole generare un *proposal*. A partire dalle immagini date in ingresso, restituisce un insieme di bounding box rappresentanti la posizione della persona nei frames desiderati. L'implementazione dello script è stata scritta in Python con la libreria OpenCV.

Motion detection La tecnica di *motion detection* utilizzata pone le sue basi su quella di *background substraction (BS)*. Dato un generico frame, rimuovendo lo sfondo della scena ottengo una *foreground mask*, ovvero un'immagine binaria contenente i pixel in corrispondenza degli oggetti che si muovono nella scena.

Le tecniche di BS calcolano la *foreground mask* sottraendo al frame corrente il *background model*, addestrato a partire da una serie di frame *statici* forniti in precedenza. La tecnica di BS utilizzata nel progetto è basata sulle *Mixture of Gaussians (MoG)*, direttamente implementate nella libreria OpenCV.

Il metodo MoG opera modellando ciascun pixel come una *mixture of Gaussians* e usa un'approssimazione per aggiornare il modello in linea. I pixel che non rispettano questa approssimazione, o non *fittano* il modello così generato, sono chiamati pixel di *foreground*.

Figura 10. Un esempio di applicazione della tecnica di BS ad un frame

Il *background model* viene inizialmente addestrato utilizzando 40 frames presi casualmente da tutta la sequenza, così da garantire una varianza più alta possibile.

L'immagine di foreground ottenuta viene processata mediante operazioni morfologiche, apertura e dilatazione, per andare a rimuovere le regioni molto piccole ed unire le regioni molto vicine tra loro. A questo punto vengono estratti i *contorni* delle regioni rimanenti definendo per ciascuna di esse una bounding box che le contiene. Le bounding box inferiori ad un una certa area prefissata (impostata a 500 pixels) vengono scartate).

Pedestrian detection Il *pedestrian detector* utilizzato si basa sull'approccio combinato di HOG e SVM. HOG, *Histogram of Oriented Gradients*, è un descrittore globale basato su istogrammi che misurano le orientazioni ed i moduli dei grandienti in una regione dell'immagine.

Il calcolo delle feature HOG si effettua in più passaggi:

- Calcola gli istogrammi in una parte dell'immagine, una *detection window* tipicamente 64×128 pixels.
- Crea blocchi di pixel 8×8 e calcola gli istorgrammi normalizzati.
- Concatena gli istogrammi ottenuti.

In Figura 11 vengono mostrate le *detections* dei due metodi distinti.

Filtro di Kalman Il *filtro di Kalman* è un filtro ricorsivo utilizzato per stimare lo stato successivo di un sistema dinamico a partire da una serie di osservazioni passate che possono essere soggette a rumore. Il filtro è utilizzato in questo elaborato per stimare la posizione successiva del soggetto in caso di assenza di informazioni dagli altri due metodi implementati.

In questo caso il filtro avrà come *stato* le coordinate del centro della bounding box. Sia X lo *stato del sistema*

La combinazione delle tecniche La predizione per ciascuna frame utilizza una combinazione delle tecniche sopra descritte. In base alla bounding box del frame precedente viene intanto selezionata una predizione per ciascuno dei

Figura 11. Rilevazione oggetti in movimento nella scena tramite *background substraction* (in mezzo) e tramite *pedestrian detection* (in basso). In alto l'immagine originale.

metodi con la condizione che sia parzialmente sovrapposta e compatibile con la previsione passata.

Scartate le bounding box non rilevante, per quelle rimanenti viene calcolato uno *score* definito come:

$$score(r_1, r_2) = \frac{intersection(r_1, r_2)}{union(r_1, r_2)} \qquad (1)$$

con r_1 e r_2 rispettivamente la bounding box del frame precedente e la bounding box predetta con una delle due tecniche (motion o pedestrian detection).

A questo punto viene selezionata la predizione con il punteggio più alto ed il filtro di Kalman viene aggiornato con il centro della bounding box selezionata.

Se non viene rilevata alcuna predizione da motion o pedestrian detection viene considerata valida quella ottenuta tramite filtro di Kalman.

4.2.2 Geometria della scena

5. Implementazione

6. Conclusioni

Conclusioni dell'elaborato

Riferimenti bibliografici

[1] F. Bartoli, G. Lisanti, L. Seidenari, S. Karaman, and A. Del Bimbo. Museumvisitors: A dataset for pedestrian and group detection, gaze estimation and behavior understanding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 19–27, 2015.

[2] F. Bartoli, L. Seidenari, G. Lisanti, S. Karaman, and A. Del Bimbo. Watts: A web annotation tool for surveillance scenarios. In *Proceedings of the 23rd ACM International Conference on Multimedia*, MM '15, pages 701–704, New York, NY, USA, 2015. ACM.

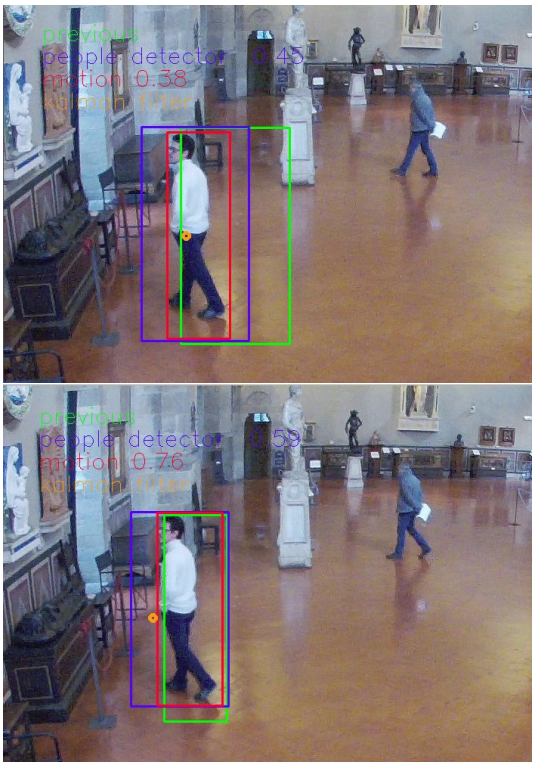

Figura 12. Esempio di esecuzione dello script con visualizzazione dei risultati parziali su due frame consecutivi. In *verde* la bounding box nella posizione precedente, in *blu* il risultato della pedestrian detection, in *rosso* il risultato del motion detector e in *arancione* il centro predetto tramite filtro di Kalman

[3] B. C. Russell, A. Torralba, K. P. Murphy, and W. T. Freeman. Labelme: A database and web-based tool for image annotation. *Int. J. Comput. Vision*, 77(1-3):157–173, May 2008.

[4] C. Vondrick, D. Patterson, and D. Ramanan. Efficiently scaling up crowdsourced video annotation. *Int. J. Comput. Vision*, 101(1):184–204, Jan. 2013.