1 OpenPose

1.1 Pose Machines

Definimos la ubicacin de la parte p, a nivel de pixeles, como $Y_p \in Z \subset \mathbf{R}^2$ donde Z es un conjunto de pares (u, v) de todas las coordenadas de pixeles en la imagen. El objetivo es predecir las localizaciones $Y = (Y_1, ..., Y_P)$ para todas las partes humanas P. Una pose machine consiste de una secuencia de predictores de multiples clases $g_t(.)$ los cuales son entrenados para predecir las ubicaciones en cada nivel de la estructura. En cada estado $t \in 1...T$ el clasificador g_t predice el nivel de confianza de asignar una posicin determinada a cada parte donde $Y_p = z, \forall z \in Z$ basandose en features extraidos de la imagen en la ubicacion z, denotados $x_z \in \mathbf{R}^d$, y mediante informacion contextual proveniente de los clasificadores vecinos al rededor de Y_p en el estado t. Un clasificador en el primer paso, $t = 1$ produce los siguientes valores de confianza:

$$g_1(x_z) \rightarrow \{b_1^p(Y_p = z)\}_{p \in \{0..P\}} \tag{1}$$

Donde $b_1^p(Y_p = z)$ es el puntaje predecido por el clasificador g_1 a la asignacion de la p^{th} parte en el primer paso en la coordenada z de la imagen. Para representar todos los valores de confianza de una parte p evaluado en todos los puntos $z = (u, v)^T$ de la imagen como $\mathbf{b}_t^p \in \mathbf{R}^{wxh}$ donde w y h son el ancho y el alto de la imagen respectivamente.

$$\mathbf{b}_t^p[u, v] = b_t^p(Y_p = z) \tag{2}$$

Por conveniencia el conjunto de los mapas de confianza de todas las partes se denota como $\mathbf{b}_t \in \mathbf{R}^{wxhx(P+1)}$).

En los siguientes estados, el clasificador predice el nivel de confianza para toda asignacion $Y_p = z$ basandose en los features de la imagen $x_z^t \in \mathbf{R}^d$ y la informacion contextual procedente de los clasificados vecinos al rededor de Y_p:

$$g_t(x_z^{'}, \psi_t(z, \mathbf{b}_{t-1})) \rightarrow \{b_t^p(Y_p = z)\}_{p \in \{0...P+1\}} \tag{3}$$

Donde $\psi_{t \geq 1}(\cdot)$ es un mapeo desde los mapas de confianza $\mathbf{b}_{t-1}$ a los context features. En casa paso las confianzas computadas proveen una estimacion cada vez mas refinada de las posiciones de las partes

1.2 Deteccion y asociacin simultanea

La arquitectura de OpenPose predice simultaneamente los confidence maps y los affinity fields. La red neuronal es dividida en dos ramas, una de las cuales preidce los confidence maps y la otra los affinity fields. Cada rama es una arquitectura de prediccion iterativa, basada en Wei et al. , en donde las predicciones se van refinando en los sucesivos estadios con supervisiones intermedias en cada uno de los estadios.

La imagen es primero analizada por una red convuncional (VGG-19), generando un conjunto de features F, los cuales son la entrada inicial para cada una de

las ramas. En el primer paso la red produce un conjunto de confidence maps $S^1 = \rho^1(F)$ y un conjunto de part affinity fields $L^1 = \phi^1(F)$ donde ρ^1 y ϕ^1 son los CNNs inference para el Paso 1. Como paso intermedio las predicciones de ambas ramas del paso previo, junto con los Features F originales de la imagen, son concatenados y usados para producir predicciones refinadas

$$\mathbf{S}^t = \rho^t(\mathbf{F}, \mathbf{S}^{t-1}, \mathbf{L}^{t-1}), \forall t \geq 2 \tag{4}$$

$$\mathbf{L}^t = \phi^t(\mathbf{F}, \mathbf{S}^{t-1}, \mathbf{L}^{t-1}), \forall t \geq 2 \tag{5}$$

Para guiar a la red para iterativamente predecir los confidence maps y los PAFS, se aplican dos funciones de perdida al final de cada estado, una para cada una de las ramas. Se utiliza una perdida L_2 entre las predicciones estimadas y los groundtruth maps y PAFs. Se le agregan pesos a las funciones de perdida para poder abordar el problema donde algunos dataset no etiquetan a toda la gente en las imagenes. Las funciones de perdida en el estado t quedan definidas como:

$$f_{\mathbf{S}}^t = \sum_{j=1}^{J} \sum_{z} \mathbf{W}(z) \cdot ||\mathbf{S}_j^t(z) - \mathbf{S}_j^*(z)||_2^2 \tag{6}$$

$$f_{\mathbf{L}}^t = \sum_{c=1}^{C} \sum_{z} \mathbf{W}(z) \cdot ||\mathbf{L}_c^t(z) - \mathbf{L}_c^*(z)||_2^2 \tag{7}$$

Donde $\mathbf{S}_c^*(z)$ es el *groundtruth* mapa de confianza, $\mathbf{L}_c^*(z)$ es el *groundtruth* part affinity field y $\mathbf{W}$ es una mascara binaria donde $\mathbf{W}(z) = 0$ cuando la anotacin no esta disponible en la coordenada z de la imagen. Esta mascara es utilizada para evitar los true positive durante el entrenamiento. El problema del vanishin gradient es abordado mediante la supervision intermedia, replenishing el gradiente perdiocamente. el objetivo es

$$f = \sum_{t=1}^{T} (f_{\mathbf{S}}^t + f_{\mathbf{L}}"t) \tag{8}$$

1.3 Confidence maps

Para evaluar f_S durante entrenamiento, se generan los groundtruth mapas de confianza $\mathbf{S}^*$ a partir de los 2D keypoints anotados. Cada mapa de confianza es una representacion 2D de la confianza the una parte del cuerpo en particular que oocurre en cada pixel de la imagen. Idealmente si solo aparece una unica persona en la imagen un solo pico deberia existir en cada mapa de confianza, en cambio, si hay multiples personas, deberia haber un pico correspondiente a cada parte visible j para cada persona k. Primero se generan los mapas de confianza individuales $\mathbf{S}_{j,k}^*$ apra cada persona k. Sea $x_{j,k} \in \mathbf{R}^2$ sea la posicion grountruth

de la parte j para la persona k. El valor en la posicion $z \in \mathbf{R}^2$ es definido en $S^*_{j,k}$ como:

$$S^*_{j,k}(z) = exp(\frac{||z - x_{j,k}||^2_2}{\theta^2}) \tag{9}$$

donde θ controla la amplitud del pico. El mapa de confianza groundtruth a ser predecido por la red, es una funcion de agfregacion de los mapas iindividuales mediante un operacion de maximo

1.4 Asociacin de Partes

Dado un conjunto de partes del cuerpo humano, se necesita una metrica que determine la asociacion entre cada par de partes, para de esta forma determinar si pertenecen a la misma persona. Una forma posible de medir el grado de asociacion es utilizando puntos intermedios entre dos partes y chequear la incidencia entre las partes detectadas como candidatas. Esta aproximacion tiene limitaciones cuando la cantidad de personas es mucha y se encuentran muy cerca unas de otras, estos puntos intermedios podrian soportar falsas asociaciones.

Para enfrentar estas limitaciones OpenPose introdujo el concepto de part affinity fields, los cuales preservan informacion sobre la ubicacion y la orientacion de la extremidad. La part affinity es un campo vectorial 2D para cada extremidad, donde por cada pixel que pertenece a una extremidad en particular, un vector 2D determina la direccion de puntos desde una parte de la extremidad hasta la otra. Cada tipo de extremidad tiene un correspondiente affinity field uniendo las dos partes asociadas.

Considerando una extremidad, sea $x_{j1,k}$ y $x_{j2,k}$ las posisiciones groundtruth de la parte j1 y j2 que conforman la extremidad c para la persona k en la imagen. Si el punto p se encuentra en la extremidad, el valor $L^*_{c,k}(p)$ es un vector unitario que apunta desde j1 a j2, para los demas puntos el vector es 0.

Para evaluar $f_{\mathbf{L}}$ durante entrenamiento definimos los groundtruth part affinity field, $\mathbf{L}^*_{c,k}$ en un punto z de la imagen

$$\mathbf{L}^*_{c,k} = \begin{cases} v & \text{si z en extremidad } c,k \\ 0 & \text{c.c.} \end{cases} \tag{10}$$

donde $v = (x_{j2,k} - x_{j1,k})/||x_{j2,k} - x_{j1,k}||_2$ es un vector unitario en la direccin de la extremidad. El conjunto de puntos que determinarn la extremidad esta definido como aquellos dentro del line segment, es decir aquellos puntos z tales que

$$0 \leq v \cdot (z - x_{j1,k}) \leq l_{c,k} \tag{11}$$

$$|v_\perp \cdot (z - x_{j1,k})| \leq \sigma_l \tag{12}$$

Donde σ_l es el ancho de la extremidad en pixeles y $l_{c,k} = ||x_{j2,k} - x_{j1,k}||_2$ el largo

2 Realidad Aumentada con Marcadores

2.1 Deteccion y asociacin simultanea

Antes de la propia deteccin del marcador el sistema, primero obtiene una *imagen de intensidad* (en escala de grises), por lo que si el formato de entrada es diferente, tal como RGB, la imagen es transformada utilizando tcnicas conocidas (Digital Image Processing, tecnica para rgb to grey)

La primer tarea necesaria para la deteccion de marcadores es encontrar los bordes de un posible marcador, para ellos se utilizan comunmente dos approachs, utilizar tecnicas de thresholding sobre imagenes para detectar marcadores en la imagen binaria, o usar algoritmos de deteccion de bordes en escala de grises (73,74)

En el caso de ARToolKit, y normalmente, utilizan mtodos de thresholdin adaptativos para que sean eficientes frente a los cambios de iluminacin (75)

Luego del thresholding el sistema tiene una imagen binaria que contiene los objetos y el fondo de la imagen original. En este paso todos los objetos detectados son posibles marcadores. Generalmente, el siguiente paso es realizar un proceso de labelling, de esta forma descartar objetos que claramente no son marcadores sea por forma o tamao. En la siguiente etapa los bordes de los posibles marcadores son delimitados y sus posiciones utilizadas en tecnicas de line fitting. Luego del line fitting el sistema realiza un nuevo chequeo sobre los candidatos analizando si posen 4 lineas rectas y 4 esquinas cada uno. Finalmente, se optimiza la localizacion de las esquinas a nivel de sub-pixels

2.2 asd

A la hora de determinar que objeto detectado es un marcador vlido y cual no, los sistemas de realidad aumentada, en su mayoria, utilizan criterios simples de rapida aceptacion de esta manera, al apuntar a aplicaciones en tiempo real, se trata de disminuir el costo computacional de la deteccion de los marcadores.

El sistema debe ignorar aquellas areas que consisten de pocos pixeles, ya que aunque en esa area exista un marcador vlido, el tamao y la distancia a la camara del mismo dificultaria el calculo de su pose con alguna fidelidad vlida. Adicionalmente utilizando este approach el sistema debe prestar atencin de no eliminar areas que pertenezcan al interior de algn marcador util y vlido

Marker pose La pose de un objeto esta determinada por su localizacion y su orientacion. La posicin se puede representar mediante tres coordenadas de traslacion y su orientacion mediante tres angulos de rotacion. Por lo tanto, una pose tiene 6 grados de libertad (6 DOF).

La pose de una camara calibrada puede ser calculada de manera unica mediante un minimo de 4 puntos (coplanar but non-collinear) (72). Por lo tanto un sistema puede calcular la pose del marcador utilizando las 4 esquinas del mismo (pdf S3)

Otro detalle a considerar a la hora de calcular la matriz de la camara, es la presencia de tres sistemas de coordenadas, el proyectado en la imagen de la camara (2D), y los sistemas 3D de la camara y el marcador.

Camera transofmration Los sistemas de coordenadas de la camara y del marcador difierente entre si por la rotacion y la trasalacion. La relacin entre ambos puede ser descripta como:

$$\begin{bmatrix} X_C \\ Y_C \\ Z_C \\ 1 \end{bmatrix} = \begin{bmatrix} R_{11} & R_{12} & R_{13} & T_1 \\ R_{21} & R_{22} & R_{23} & T_2 \\ R_{31} & R_{32} & R_{33} & T_3 \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} X_M \\ Y_M \\ Z_M \\ 1 \end{bmatrix} = T_{CM} * \begin{bmatrix} X_M \\ Y_M \\ Z_M \\ 1 \end{bmatrix}$$

Donde T_{CM} es la matriz de pose o la matriz de transformacin de la camara, la cual se utiliza para determinar la posicin del objeto a aumentar con respecto en el sistemas de coordenadas de la camara. Los valores de T_i corresponden al vector de traslacion, mientras que los nueve restantes R_{ij} son parametros obtenidos a partir de las 3 coordenadas de rotacin. Esta operacin se realiza en cada frame donde un marcador es detectado.

La relacin entre las coordenadas de la imagen de la camara y y las coordenadas de la camara esta definida por:

$$\begin{bmatrix} hX_1 \\ hY_1 \\ h \end{bmatrix} = \begin{bmatrix} P_{11} & P_{12} & P_{13} & 0 \\ 0 & P_{22} & P_{23} & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} X_C \\ Y_C \\ Z_C \\ 1 \end{bmatrix} = P \begin{bmatrix} X_C \\ Y_C \\ Z_C \\ 1 \end{bmatrix}$$

donde P se denomina *intrinsic camera parameters*, la cual es camara dependiente. En el caso de ARToolKit esta misma esta definida como:

$$\begin{bmatrix} s_x f & 0 & x_0 & 0 \\ 0 & s_y f & y_0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix}$$

donde f es la longitud focal de la camara, s_x el factor de escala en el eje x, s_y el factor de escala en el eje y y por ltimo (x_0, y_0) es la posicion donde el eje z del sistema de coordenadas de la camara *frame passes*

2.3 Pose calculation

Dada las caracteristicas de las camaras modernas,podemos asumir que la distorsion puede ser separada del modelo de la camara. Aquellos puntos que pertenecen a las coordenadas sin distorsion

2.4 asd

las coordenadas reales de las esquinas son conocidas. El sistema poee ocho ecuaciones, una para cada una de las coordenadas que componen los cuatro puntos

de las esquinas, y 6 parametros libres. Esto posibilita la estimacion de la matriz de transformacin.

Un approach comunmente usado es utiliza metodos no iterativos para obtener una estimacion inicial de la pose, por ejemplio DTL direct linear transformation, y luego utilizar una optimizacion iterativa para calcular la pose exacta.

Se reproyecta el mundo X en el plano de la imagen, utilizando la matriz estimada $M^{'}$, definimos el punto reproyectado $x^{'}$

$$x^{'} = M^{'}X \tag{13}$$

Se puede resolver la matriz de tranformacion encontrando una matriz que miniza el error de proyecccion $||x - x^{'}||$

$$err = \frac{1}{4}\sum i = 14 \tag{14}$$

Este es un problema de estimacion no-lineal y el sistema puede resolverse analogamente a la calibracion de la camera utilizando el metodo iterativo Levenberg

La optimizacion de lerror de reproyeccion es un metodo rapido y es utilizado por ARToolKit. En el modo de trackeo continuo, ARToolKit combina esta tecnicae con los resultados de tracking del frame anterior [83]

Template matvhing Template markers son marcados en blanco y negro que contienen una imagen simple dentro de bordes negros. El sistema de deteccion identifica los marcados comparandolos con markers template, son ejemplos de marcadores, los cuales tienen un nombre o id unico que los identifica. De esta manera se elija la mayor comparacion.

La identificacion de marcadores es una version simplificada del problema de template matching, ya que en la deteccion de marcadore el area de matching eta definida, mientras que en el problema general, la ubicacion, tamao y orientacion de la matching area no son conocdas.

En template matching, el marcador detectado es unwarped utilizando la pose de la camara calculada, escalando al tamao del template el marcador y comparando en cuatro posicione diferentes el marcador con los distintos templates. El template que da el mayor valor de similitud (el menor de disimilitud) e el marcador correcto, adicionalmente obteniendo la orientacion correcta.

Si todos los valores de similitud son menores al trheshold el sistema descarta el marcador.

En lugar de unrwrapping todo el marcador, el sistema puede proyectar los centros de las cells en el template a las coordenadas de la magen utilizando la pose de la camara calculada. Luego los valores de los pixeles se puede obtener directamente de la imagen en escala de gris o threshold. El valor de los pixeles se puede coniderar como el valor del pixel mas cercano, el promerio de los N vecinos.

Antes del template matching, generalmente, una imagen en escala de gris es normalizada de modo que aquellas zonas oscuras quedan en negro y las claras en blanco.

el valor de similitud puede basarse en SSD (suma de diferencias de cuadrados) or cross-correlation

2.5 Context features

Para obtener correlaciones epaciale entre los niveles de confianza de cada parte con respecto a sus vecinos, se describen dos tipos de factores, "context" feature maps denotados por ψ_1 y ψ_2.

Context patch features.