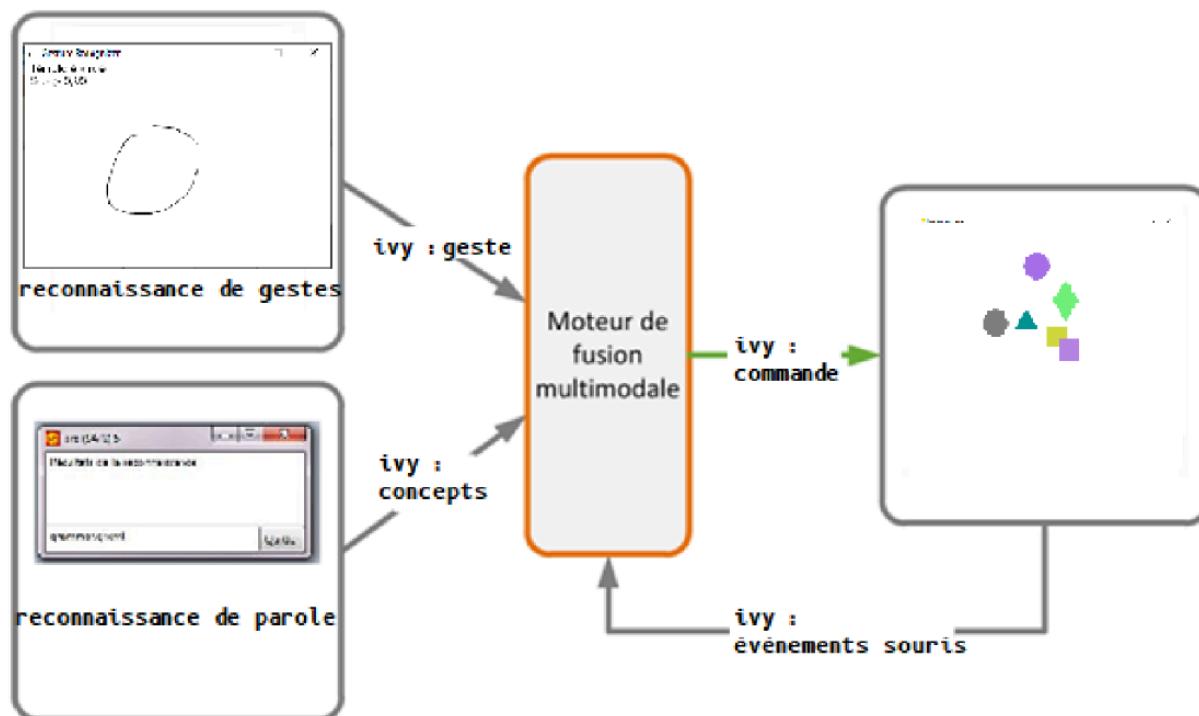


MOTEUR DE FUSION MULTIMODAL

– COMPTE RENDU –



Ecrit et réalisé par : BAFFOGNE Clara et BLAYES Hugo

Encadré par : TRUILLET Clément

SOMMAIRE

1 Introduction et Contexte	3
2 Gestion de la parole	3
3 Gestion du geste – par dessin	4
4 Gestion du geste – par hand tracking	5
5 Souris	7
6 Moteur multimodal	7
7 Interface	7
8 Lancement	8
9 Conclusion	8

1 Introduction et Contexte

L'objectif principal de ce projet est le développement d'un moteur de fusion multimodal permettant à l'utilisateur de créer des formes graphiques en choisissant une couleur spécifique. Ce moteur intègre la reconnaissance de la parole et la reconnaissance de geste. La position de la souris permet de positionner la forme à l'emplacement précis du curseur dans la fenêtre graphique.

Il existe quatre formes possibles :

- Carré
- Rectangle
- Triangle
- Cercle

Nous avons implémenté trois couleurs :

- Rouge
- Bleu
- Vert

Le moteur multimodal regroupe la reconnaissance de parole, la reconnaissance de geste par dessin mais également par le suivi des mouvements de la main (hand tracking) pour effectuer les actions demandées par l'utilisateur à la position de la souris.

2 Gestion de la parole

Pour gérer la reconnaissance vocale, nous utilisons la librairie Speech Recognition. Cette dernière repose sur les Modèles Cachés de Markov pour convertir les sons captés par le microphone en coefficients cepstraux. Ces coefficients sont ensuite transformés en phonèmes, qui sont traduits en phrases cohérentes grâce à une vaste base de données. Plus précisément, les phonèmes sont envoyés à une API Google, utilisée notamment pour des outils comme Google Home, qui renvoie la phrase jugée la plus probable.

Un des atouts majeurs de cette librairie est son filtre anti-bruit, qui améliore la précision même dans un environnement sonore perturbé. Nous avons ensuite intégré cette librairie avec un module capable de transmettre les informations reçues à un serveur Ivy, ce qui permet de centraliser les données captées par le microphone de l'utilisateur.

La reconnaissance de la parole consiste à dire à l'oral une action suivie de la couleur. En parallèle, l'utilisateur doit positionner la souris dans la fenêtre pour indiquer au système où il souhaite réaliser l'action.

Dans le cas de la reconnaissance de la parole, les quatre actions possibles sont :

- Crée : reconnu par les mots « créer » et « dessiner ». Cette action permet de créer la forme demandée par l'utilisateur.
- Modifier : modifie la couleur de la forme indiquée par l'utilisateur.
- Effacer : reconnu par les mots « supprimer » et « effacer ». Cette action supprime la forme souhaitée.
- Déplacer : permet de déplacer la forme souhaitée.

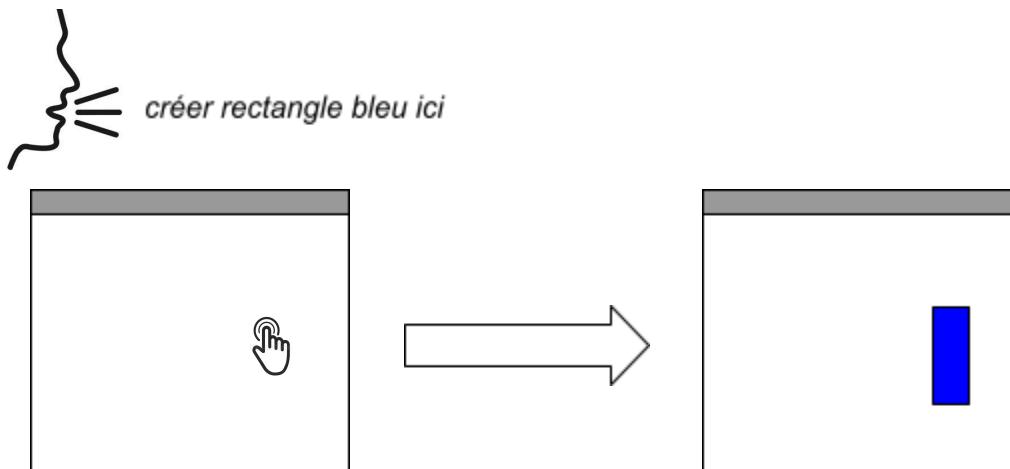


Figure 1 : Exemple de création de forme en reconnaissance de parole

3 Gestion du geste – par dessin

Pour la reconnaissance de geste, nous avons utilisé l'algorithme du One Dollar Recognizer. Cet algorithme est particulièrement adapté à la reconnaissance de formes simples dessinées par l'utilisateur.

Nous enregistrons les dessins des différentes formes afin qu'elles soient reconnues par le système. L'utilisateur peut dessiner les formes directement sur l'interface à la position de la souris.



Figure 2 : Geste triangle et cercle

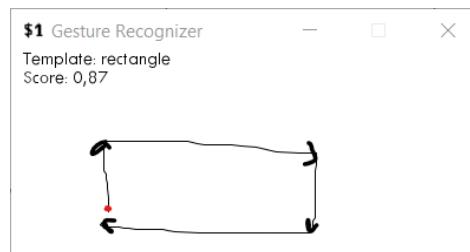


Figure 3 : Geste rectangle

Toutes les formes doivent être tracées dans le sens des aiguilles d'une montre. Pour le rectangle, il est important de commencer le dessin soit par le coin inférieur gauche, soit par le coin supérieur droit afin qu'il soit correctement reconnu.

4 Gestion du geste – par hand tracking

Pour la gestion du geste via hand tracking, dans un premier temps nous avons utilisé la librairie mediapipe. Cette librairie créée par Google permet de détecter, entre autres, si une main est présente sur une image si c'est le cas, on peut récupérer la position de chaque phalange de chaque doigt. A l'instar du One dollar, on peut stocker les positions puis utiliser un algorithme de reconnaissance permettant de détecter si notre liste de position forme une figure.



*Figure 4 : Dessiner forme
hand tracking*



*Figure 5 : Envoyer forme à
l'interface*

Les figures 4 et 5 illustrent les gestes nécessaires pour créer une forme dans l'interface. D'abord, nous dessinons la forme en utilisant l'index (voir fig. 4), puis nous validons la forme en ouvrant la paume de la main (voir fig. 5).



Figure 6 : Prendre la forme pour la déplacer



Figure 7 : Relâcher la forme

Les figures 6 et 7 illustrent les gestes à réaliser pour déplacer une forme dans l'interface. Les formes affichées dans l'interface sont redessinées dans la fenêtre de la caméra, ce qui facilite leur manipulation. Ces gestes suivent le principe du "pick and place", c'est-à-dire que l'on saisit la forme en pinçant l'index et le pouce (ref. fig 6), puis on la relâche à l'endroit souhaité (ref. fig 7).



Figure 8 : Nettoyer l'affichage caméra

La figure 8 présente la procédure pour réinitialiser l'affichage de la fenêtre de la caméra. Pour ce faire, il suffit de lever simultanément l'index et le majeur.

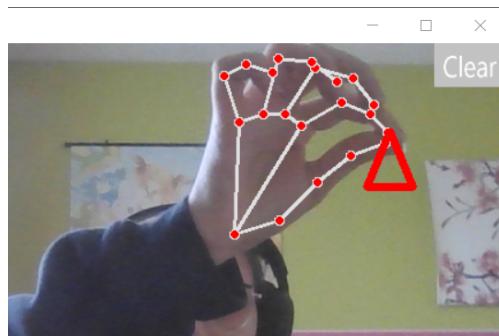


Figure 9 : Suppression forme

Pour supprimer, l'utilisateur peut déplacer la forme vers la zone "Clear" qui se situe en haut à droite de la fenêtre de la caméra.

5 Souris

La souris fonctionne en complément de la reconnaissance des gestes et de la parole. Elle permet d'effectuer une action à l'endroit souhaité dans la fenêtre. La position de la souris est associée aux termes vocaux tels que "ici", "là", ou "cette forme", qui peuvent être prononcés oralement.

Lors de la création (resp. déplacement) d'une forme, l'utilisateur place la souris dans l'interface et prononce "créer forme ici" ou "dessine" (resp . "déplace"). La forme est alors générée aux coordonnées de la souris.

Dans les actions de modification ou de suppression, l'utilisateur doit se positionner sur la forme afin d'indiquer au système sur quelle forme intervenir.

6 Moteur multimodal

Le moteur multimodal assure la connexion entre les diverses entrées multimodales et notre interface. Pour combiner ces entrées, il est nécessaire de réunir les résultats de chaque agent (OneDollar, PyVocal, PyMove et ClickSouris). Pour ce faire, nous avons opté pour l'utilisation d'Ivy. Ivy offre la possibilité de créer un serveur permettant l'envoi des informations. Un de ses atouts est qu'il facilite la programmation hybride en intégrant plusieurs langages, tels que Python et Processing-Java. Ensuite, notre moteur multimodal lit et traite chaque message reçu sur le serveur de manière continue. Ce traitement en temps réel apporte une grande flexibilité à notre système.

7 Interface

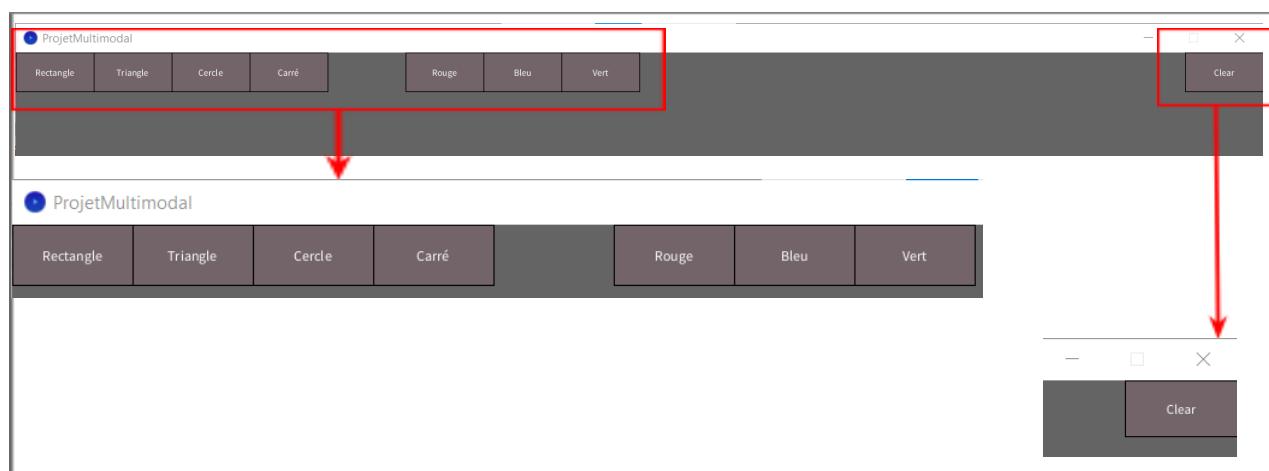


Figure 10 : Interface

Sur la partie gauche de l'interface, se trouvent tous les boutons nécessaires à la création des formes, avec la forme à gauche et la couleur à droite. Sur la droite de l'interface, le bouton “Clear” permet de supprimer toutes les formes dessinées.

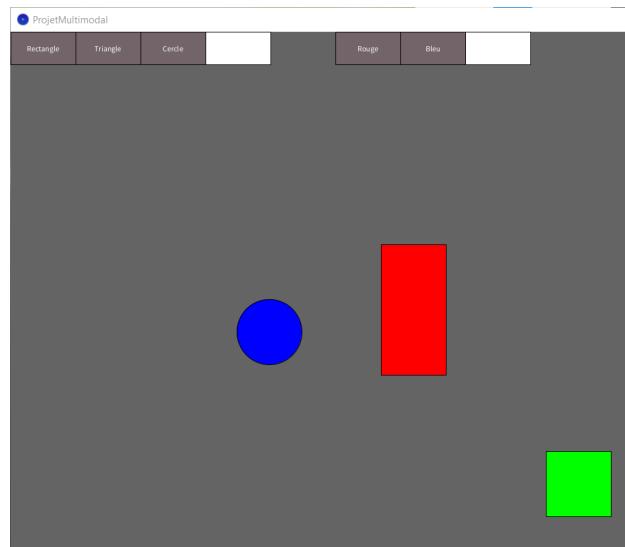


Figure 11 : Création forme

En cliquant sur une forme et sur une couleur, et enfin en cliquant à un endroit quelconque dans l'interface, la forme sera créée à cet emplacement.

8 Lancement

Pour démarrer le projet, ouvrez le fichier “ProjetMultimodal.pde”. Cela lancera les modules de reconnaissance vocale et de hand tracking. À ce moment, une fenêtre grise représentant l'interface apparaîtra, accompagnée de deux terminaux : l'un pour la reconnaissance vocale et l'autre pour le hand tracking, ce dernier ouvrant également une fenêtre de caméra.

Pour activer la reconnaissance de gestes, ouvrez le fichier “OneDollarIvy.pde” situé dans le dossier “OneDollarIvy”. Une petite fenêtre blanche s'affichera. Appuyez sur la touche “I” pour commencer la reconnaissance des gestes.

9 Conclusion

Ce moteur multimodal permet de combiner la reconnaissance vocale, la reconnaissance de gestes (dessin et hand tracking), et l'utilisation de la souris pour interagir avec une interface graphique. Grâce à cette approche, il est possible de créer, modifier, déplacer ou supprimer des formes, tout en utilisant différentes modalités en parallèle.