The Kinect provides you with the skeletons it's tracking, you have to do the rest. Basically you need to create a definition for each gesture you want, and run that against the skeletons every time the SkeletonFrameReady event is fired. This isn't easy.

**Defining Gestures**

Defining the gestures can be surprisingly difficult. The simplest (easiest) gestures are ones that happen at a single point in time, and therefore don't rely on past locations of the limbs. For example, if you want to detect when the user has their hand raised above their head, this can be checked on every individual frame. More complicated gestures need to take a period of time into account. For your waving gesture, you won't be able to tell from a single frame whether a person is waving or just holding their hand up in front of them.

So now you need to be able to store relevant information from the past, but what information is relevant? Should you keep a store of the last 30 frames and run an algorithm against that? 30 frames only gets you a second's worth of information.. perhaps 60 frames? Or for your 5 seconds, 300 frames? Humans don't move that fast, so maybe you could use every fifth frame, which would bring your 5 seconds back down to 60 frames. A better idea would be to pick and choose the relevant information out of the frames. For a waving gesture the hand's current velocity, how long it's been moving, how far it's moved, etc. could all be useful information.

After you've figured out how to get and store all the information pertaining to your gesture, how do you turn those numbers into a definition? Waving could require a certain minimum speed, or a direction (left/right instead of up/down), or a duration. However, this duration isn't the 5 second duration you're interested in. This duration is the absolute minimum required to assume that the user is waving. As mentioned above, you can't determine a wave from one frame. You shouldn't determine a wave from 2, or 3, or 5, because that's just not enough time. If my hand twitches for a fraction of a second, would you consider that a wave? There's probably a sweet spot where most people would agree that a left to right motion constitutes a wave, but I certainly don't know it well enough to define it in an algorithm.

There's another problem with requiring a user to do a certain gesture for a period of time. Chances are, not every frame in that five seconds will appear to be a wave, regardless of how well you write the definition. Where as you can easily determine if someone held their hand over their head for five seconds (because it can be determined on a single frame basis), it's much harder to do that for complicated gestures. And while waving isn't that complicated, it still shows this problem. As your hand changes direction at either side of a wave, it stops moving for a fraction of a second. Are you still waving then? If you answered yes, wave more slowly so you pause a little more at either side. Would that pause still be considered a wave? Chances are, at some point in that five second gesture, the definition will fail to detect a wave. So now you need to take into account a leniency for the gesture duration.. if the waving gesture occurred for 95% of the last five seconds, is that good enough? 90%? 80%?

The point I'm trying to make here is there's no easy way to do gesture recognition. You have to think through the gesture and determine some kind of definition that will turn a bunch of joint positions (the skeleton data) into a gesture. You'll need to keep track of relevant data from past frames, but realize that the gesture definition likely won't be perfect.

**Consider the Users**

So now that I've said why the five second wave would be difficult to detect, allow me to at least give my thoughts on how to do it: don't. You shouldn't force users to repeat a motion based gesture for a set period of time (the five second wave). It is surprisingly tiring and just not what people expect/want from computers. Point and click is instantaneous; as soon as we click, we expect a response. No one wants to have to hold a click down for five seconds before they can open Minesweeper. Repeating a gesture over a period of time is okay if it's continually executing some action, like using a gesture to cycle through a list - the user will understand that they must continue doing the gesture to move farther through the list. This even makes the gesture easier to detect, because instead of needing information for the last 5 seconds, you just need enough information to know if the user is doing the gesture right now.

If you want the user to hold a gesture for a set amount of time, make it a stationary gesture (holding your hand at some position for x seconds is a lot easier than waving). It's also a very good idea to give some visual feedback, to say that the timer has started. If a user screws up the gesture (wrong hand, wrong place, etc) and ends up standing there for 5 or 10 seconds waiting for something to happen, they won't be happy, but that's not really part of this question.

**Starting with Kinect Gestures**

Start small.. really small. First, make sure you know your way around the SkeletonData class. There are 20 joints tracked on each skeleton, and they each have a TrackingState. This tracking state will show whether the Kinect can actually see the joint (Tracked), if it is figuring out the joint's position based on the rest of the skeleton (Inferred), or if it has entirely abandoned trying to find the joint (NotTracked). These states are important. You don't want to think the user is standing on one leg simply because the Kinect doesn't see the other leg and is reporting a bogus position for it. Each joint has a position, which is how you know where the user is standing.. piece by piece. Become familiar with the coordinate system.

After you know the basics of how the skeleton data is reported, try for some simple gestures. Print a message to the screen when the user raises a hand above their head. This only requires comparing each hand to the Head joint and seeing if either hand is higher than the head in the coordinate plane. After you get that working, move up to something more complicated. I'd suggest trying a swiping motion (hand in front of body, moves either right to left or left to right some minimum distance). This requires information from past frames, so you'll have to think through what information to store. If you can get that working, you could try string a series of swiping gestures in a small amount of time and interpreting that as a wave.

tl;dr: Gestures are hard. Start small, build your way up. Don't make users do repetitive motions for a single action, it's tiring and annoying. Include visual feedback for duration based gestures. Read the rest of this post.